Trust and Cheating[*]

Jeffrey V. Butler
EIEF

Paola Giuliano
UCLA, NBER, CEPR and IZA

Luigi Guiso
EIEF and CEPR

This version: September 2013

Abstract

When we take a cab we may feel cheated if the driver takes an unnecessarily long route despite the lack of a contract or promise to take the shortest possible path. Is the behavior of the driver affected by beliefs about our cheating notions, and where do his beliefs come from? For that matter, where do our cheating notions come from, and how do they color our own decisions? We address these questions in the context of a trust game by asking participants directly about their personal notions of cheating. We find that both parties to a trust exchange have personal notions of what constitutes cheating; that these notions have a bimodal distribution; and that cheating notions are determined by parentally-transmitted values. We also document that cheating notions substantially affect decisions on both sides of the trust exchange. We finally link our results to guilt aversion models and provide a more general view on the type of preferences that could explain receivers' cheating decisions.

JEL Classification: A1, A12, D1, O15, Z1
Keywords: Trust, trustworthiness, social norms, culture, cheating

[*] We thank Martin Dufwenberg, Uri Gneezy and Roland Bénabou as well as seminar participants at EIEF and conference participants at the Sciences Po/IZA Workshop on Trust, Civic Spirit and Economic Performance, the 2011 Florence Workshop on Behavioural and Experimental Economics and the 2011 SITE Summer workshop at Stanford University for many helpful comments.

1 Introduction

When taking a cab we may expect the driver to use a reasonably short route even if neither we nor the driver make explicit mention of it. Despite the lack of explicit promise, we may still feel cheated if the taxi driver takes an unnecessarily long route. Similarly, when we ask for financial advice the advisor does not typically spell out that he will act solely in our best interest, but we may still judge cheating according to this metric. When we book a vacation through a travel agent, search for the best medical insurance at a broker or take our cars to a mechanic, we may act on implicit notions of what the behavior of the travel agent, broker or mechanic *should be*, perhaps feeling cheated or let down when behavior fails to live up to these standards.

Situations like these come up frequently in our daily economic lives: opportunities for mutually beneficial exchanges where complete contracts, agreements or credible communication about what is expected from each side of the exchange are either impossible or infeasible. Considering only our first example above, over $600,000$ taxi rides are taken daily in New York city alone constituting about \$1 billion in fares paid per year.[1] And New York is not alone: about one *million* people use taxis every day in Hong Kong,[2] while a staggering three to four million taxi rides are taken every day in Lima, Peru (Castillo, et al., 2012). Our second example—financial advice from professionals—is also pervasive. According to a broad survey of retail investors in Germany, more than 80 percent of investors consult a financial advisor. Overall, in the UK 91% of intermediary mortgage sales are "with advice" (Chater, Huck and Inderst, 2010). In the US, 73% of all retail investors consult a financial advisor before purchasing shares (Hung, *et al.*, 2008).[3] Given their ubiquity, understanding precisely what drives behavior in such trust-based exchange opportunities is an important undertaking.

In this paper, we focus on one intuitively plausible yet under-explored determinant of behavior on both sides of such exchange opportunities: individuals' personal, subjective, notions of what constitutes cheating. While individuals may hold widely divergent views on what constitutes cheating and this cheating notion heterogeneity may, in turn, translate into heterogeneous behavior, economists know virtually nothing about the relationship between

[1] http://en.wikipedia.org/wiki/Transportation_in_New_York_City.

[2] http://www.gov.hk/en/about/abouthk/factsheets/docs/transport.pdf

[3] See also Inderst and Ottaviani (2012) for a general review on financial advice.

personal cheating notions and behavior in trust-based exchange opportunities. For instance, in the massive body of experimental trust game literature researchers typically *assume* that both involved parties will define cheating according to a single, shared, notion.[4] While this methodology has proven useful in showing that pecuniary concerns alone fail to account for a significant portion of exchange behavior, its ability to provide a detailed understanding of how idiosyncratic cheating notions translate into behavior is obviously limited.

We investigate the role of the cheating notion in the context of a trust game (Berg, Dickhaut and McCabe, 1995), a two-player sequential moves game of perfect information. In this game, the sender moves first by deciding whether to send some, all or none of a fixed endowment to a co-player, the receiver. Any amount sent is increased by the experimenter before being allocated to the receiver, who then decides whether to return some, all or none of this (increased) amount to the sender. While highly stylized, the trust game is an appropriate context because it captures the essential nature of our motivating examples: a pareto-improving exchange is possible, but comes with the risk of opportunistic counterparty behavior which cannot be eliminated through pre-play promises or contracts.

The timing of our experiment is as follows: first, we have participants play a slightly-modified trust game; after playing the trust game, we ask participants directly about their personal, subjective, cheating notions;[5] finally, we elicit participants' beliefs about others' cheating notions and behavior, as well as their first- and second-order beliefs about others' behavior and expectations. We complement the data from our trust game experiment with data on the values our participants' parents emphasized during their upbringing. These values data were collected for a previously-conducted, unrelated, experiment that took place *from twenty days to sixty days prior* to the trust game experiment. We use these data to investigate one potential source of stable heterogeneity in cheating notions: cultural transmission.

In this specific setting we test several hypotheses. At the most basic level, we test whether such situations do indeed engender implicit cheating notions. Whether or not this

[4] For example, in a seminal work in this vein Berg, Dickhaut and McCabe (1995) explicitly posit that trustors will feel cheated by a negative return on their trust-investment. This often unstated assumption continues to pervade the trust literature: the outcome chosen to highlight the existence of aversion to "betrayal," or what we would call cheating, is one that falls just below yielding a positive return on investment (see, e.g., Bohnet and Zeckhauser, 2004).

[5] We realize that asking about cheating notions directly gives rise to concerns about priming. We check for the robustness of directly-revealed cheating notions in additional robustness sessions where cheating notions are elicited indirectly in a way that reduces the likelihood of priming effects.

will be the case is not *a priori* obvious: cab drivers, mechanics and financial advisors may very well choose to ignore or downplay the possibility that their customers could ever feel cheated in order to reconcile opportunistic behavior with a positive self-image.[6] Secondly, conditional on an affirmative answer to our first question, we test the hypothesis that these implicit cheating notions have an impact in determining behavior in a trust-based exchange situation.

We find that the vast majority of participants articulate a cheating notion even when they can easily refrain from doing so, suggesting they are genuine. We document these notions, showing they are roughly bimodal: many participants define cheating by a positive return on investment rule, as assumed but not tested by Berg, Dickhaut and McCabe (1995); while, contrary to the assumptions of much of the trust game literature, a sizable minority of senders (around 30% of participants) define cheating by a more demanding notion requiring fully half of their co-players' total earnings in order to not feel cheated.[7] We also show that this heterogeneity in cheating notions carries over to beliefs about others' cheating notions. On our second point, we find that the notion of cheating strongly affects behavior on both sides of the potential exchange.

Our paper contributes to the literature in several ways. First and foremost, we provide the first direct evidence on the relationship between personal cheating notions and individual behavior in trust-based exchange opportunities.

Second, we provide evidence of a substantial role for culturally transmitted values in the formation of cheating notions and related beliefs. Our data are consistent with the plausible notion that parentally instilled values exert a substantial impact on how individuals define cheating and that own cheating notions shape beliefs about others' cheating notions. Together, these patterns suggest there may be a substantial temporally stable component of personal cheating notions, adding to their predictive value.

The third contribution is the investigation of how cheating notion beliefs constrain the behavior of the *entrusted*. We find that the behavior of individuals who refrain from intentional cheating moves in one-to-one correspondence with their beliefs about others' cheating notions. On the other side of the exchange, we document a significant relationship

[6] For evidence that individuals choose their beliefs to avoid cognitive dissonance, we refer the interested reader to the discussion in Akerlof and Dickens (1982).

[7] Because of the way we modified the trust game, this latter rule can be distinguished from previously documented fairness rules such as equal surplus division. For details, see the experimental design section.

between expected cheating and *how much* individuals trust.

Fourth, we provide empirical content to theoretical models of guilt aversion (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007). In particular, we show that cheating notions and beliefs about others' cheating notions exhibit a close empirical relationship with the relevant first- and second-order beliefs. In addition, we show that guilt arising from cheating constrains receivers' behavior.

More generally, our results contribute to the debate over how non-pecuniary preferences affect behavior and where these preferences come from. Receivers in the trust game face a stark trade-off between their pecuniary preferences and moral behavior. Our finding that receivers' behavior is affected by their beliefs about what constitutes cheating lends support to the view put forward by Gneezy (2005): moral preferences are affected by the magnitude of damage that immorality inflicts on others.[8] However, because our experiment involves a game with neither communication nor unambiguous moral standards, and hence no literal lying nor deception, we extend Gneezy's findings by showing that the moral forces at work operate outside of the specific context of deception.

Our paper is also related to a nascent literature examining directly the relationship between behavior and social norms that is exemplified by Krupka and Weber (2013). Similar to our study, the aim of this body of research is to complement the copious indirect evidence that social norms affect behavior by directly eliciting social norms and relating the elicited social norms to observed behavior. Our study differs from this vein of research, however, in that we focus on personal cheating notions which may vary widely across individuals and require no tacit or explicit agreement about what is cheating and what is not. In stark contrast, social norms by definition require a "...general social agreement that some actions are more or less socially appropriate" (*ibid*).

Finally, our paper relates to the huge literature investigating behavior in the trust game.[9] The bulk of this vast literature focuses on what drives senders' behavior—interpreting the amount senders send as trust, whence the moniker "the trust game" comes. What, precisely, senders are trusting receivers to do is typically left unspecified, but a common assumption—made explicitly in Berg, Dick and McCabe (1995) and implicitly in much subsequent work

[8] Many popular and intuitive models of moral preferences are inconsistent with this pattern in behavior. For an elaboration of the inconsistencies, see Gneezy (2005).

[9] The trust game literature is too large and spans too many disciplines to be summarized here, but for an excellent review see Camerer (2003) and the references therein.

(e.g., Bohnet and Zeckhauser, 2004)—is that senders are trusting that receivers will send back at least as much money as they sent. To the best of our knowledge, this assumption has never been tested directly. Differently from most of this literature, rather than assuming a particular cheating notion is operative, we aim to document participants' personal cheating notions and beliefs about others' cheating notions and study their role in determining the behavior of *both* receivers and senders. In doing so, we can shed empirical light on the unresolved question of what it is that senders are trusting receivers to do, what receivers believe senders are expecting of them, and the determinants of receivers' behavior.

The remainder of the paper proceeds as follows: Section 2 details the experimental design; Section 3 presents the results. Section 4 links our paper to models of guilt aversion theory and provides a more general discussion of our findings together with a simple analytical framework to interpret them. The final section concludes and suggests avenues for future research. Additional experimental treatments, analyses conducted to address the robustness of elicited cheating notions and a comparison between behavior in our main experiment—conducted on-line—and a smaller study conducted in a more traditional laboratory environment can be found in Appendix I. Appendix II provides instructions for our main experiment.

2 Experimental Design

A total of 428 individuals participated in our main study, all of whom were students in Rome, Italy, enrolled at one of two universities: LUISS Guido Carli University or the University of Rome, La Sapienza. All sessions were conducted on-line.

The experiment consisted of three phases. First, participants played a slightly modified trust game. Responses were collected using the strategy method: participants submitted their complete contingent strategies for both the sender and receiver roles before knowing which role they would be assigned. After participants submitted their trust game strategies, we asked them about their personal cheating notions. Finally, we elicited participants' beliefs about others' behavior and others' cheating notions in an incentive compatible manner. During each of these three phases, participants were unaware of the existence of any of the subsequent phases.

2.1 Our slightly-modified trust game

Our trust game is standard in most respects: it is a two-player sequential moves game of perfect information involving a sender and a receiver. The sender moves first by deciding whether to send some, all or none of a fixed endowment to the receiver. Any amount sent is increased by the experimenter before being allocated to the receiver, who then decides whether to return some, all or none of this (increased) amount to the sender. Pairings are random and anonymous.

Our trust game differs, however, in two important ways from the canonical trust game. First of all, we implement an unequal endowment design—senders (receivers) are endowed with 10.50 euros (0 euros). Secondly, while most trust games use a linear function to transform the amount sent into the amount received—typically, if a sender sends s, the receiver receives $f(s) = 3s$—we implement a concave trust production function. In our trust game, when a sender sends s euros, the receiver receives $8\sqrt{s}$ euros.[10] Both of these modifications will allow us to distinguish among various *a priori* likely cheating notions. For example, a fairness notion that says "I am entitled to half of the surplus created from my actions" coincides with an egalitarian fairness notion ("everybody's final money outcome should be the same") in the standard trust game with equal endowments, but not in our unequal endowment setting. A concave production function will allow us to further distinguish among various common fairness rules which roughly coincide when using a linear function.[11]

An added benefit of using a concave production function is to provide a relatively smooth relationship between behavior and beliefs at the individual level, a feature which will prove useful when we examine the "intensive margin" of trust: how much to send conditional on sending something.[12] Aiding our identification of this intensive margin is one additional,

[10] In order to allow us to use the strategy method, we restrict the sender's action set to include only integer amounts. The possible amounts a receiver could receive are $f(0) = 0, f(1) = 8.05, f(2) = 11.30, f(3) = 13.85, f(4) = 16.05, f(5) = 17.90, f(6) = 19.60, f(7) = 21.20, f(8) = 22.65, f(9) = 24.05, f(10) = 25.30$. This trust production function deviates slightly from $f(S) = 8\sqrt{s}$ in order to produce relatively simple values (multiples of 5 cents) while, at the same time, maintaining concavity and surplus creation. Surplus creation is a central feature of the trust game and refers to the fact that each additional dollar sent always produces more than one dollar in receiver earnings. Participants were presented the trust production function in table format to facilitate comprehension.

[11] For example, consider two possible cheating notions conditional on sending one euro: a positive return on investment notion; or an equal share of created surplus notion. Irrespective of the trust production function the former notion entails receivers returning 1 euro. The latter notion entails receivers returning $\frac{f(s)}{2}$, which is 1.50 euros when $f(s) = 3s$ but $\frac{8.05}{2} = 4.025$ euros using our concave trust production function. Consequently, these two notions would differ by only 0.50 euros using the standard trust production function, while, in stark contrast, our concave function separates these two cheating notions by just over 3 euros.

[12] For instance, if senders have standard risk-neutral preferences a linear trust production function often

more subtle, feature of our design. We introduce a small (0.50 euro) fixed sending fee in some sessions: in "high fee" sessions, senders who choose to send a strictly positive amount incur the fee whereas senders who chose to send nothing do not; in the remaining "low fee" sessions, senders never incur a sending fee. This provides exogenous variation in the cost of sending something versus nothing—the extensive margin of trust—which will allow us to formally model and estimate the intensive and extensive margins of trust separately.

Senders' feasible actions consisted of sending any whole-euro amount, including 0. Conditional on receiving $f(s) > 0$ euros, receivers' feasible actions were $\{0.00, 0.01, \ldots, f(s)\}$. We employed the strategy method to collect participants' trust game decisions: before discovering whether they would play the role of sender or receiver, participants submitted a complete contingent strategy for each role. Each participant specified how much they would send in the role of sender and, for each possible amount they could receive in the role of receiver, how much they would return. The order in which participants submitted their strategies—whether first for sender, then for receiver or first for receiver and then for sender—was randomized. Additionally, to bridge the gap between the strategy method and the direct response method and to attempt to make each receiver's decision feel as real as possible, participants' receiver strategies were elicited with a series of ten separate screens. Each of these ten screens asked only one question: "if the sender sends s euros and you therefore receive $f(s)$ euros, how much will you return?"[13]

2.2 The cheating notion questions

After participants submitted their complete contingent trust game strategies, we asked them to specify their personal definitions of cheating from the perspective of the sender. For each

implies corner solutions: send the entire endowment if the expected net return from trusting is positive, or nothing if the net return is negative; if the expected return is zero, then all send amounts are optimal. In contrast, our concave production function provides such senders unique internal optimal send amounts that vary continuously with expected return over a wide range of beliefs. In this sense our concave function may provide a more realistic portrait of trusting behavior outside of the lab with stakes large enough for risk aversion to matter. Consequently, an additional justification for using a concave trust production function is to induce risk-averse preferences (Smith, 1976).

[13] For each separate screen, s was replaced with exactly one of the 10 possible amounts a sender could send ($s \in \{1, \ldots, 10\}$), and $f(s)$ was replaced with the corresponding value from the trust production function ($f(s) \in \{8.05, \ldots, 25.30\}$). The order in which receivers faced their ten separate decisions was randomly predetermined but the same for all participants. This maintains comparability across observations without inducing undue consistency in receiver strategies that might arise from, e.g., facing a monotonically increasing or decreasing sequence of send amounts. The order used was $S = 7, 4, 8, 3, 9, 10, 2, 1, 6, 5$.

possible strictly positive send amount, $s \in \{1, 2, \ldots, 10\}$, participants were asked:[14]

> "If you are assigned the role of A [sender] what is the minimum amount you would need to receive back from player B [receiver] in order to not feel cheated? . . . If you were to send s euros and B were to therefore receive $f(s)$ euros, you would need back how many euros?"

To respond, participants could either insert a number between 0.00 and $f(s)$ or refrain from specifying a cheating notion by selecting one of two options: "this has nothing to do with cheating;" or "I don't know." Leaving the question blank was also allowed, but not explicitly mentioned as an option.[15]

One obvious concern with asking about cheating notions in this manner is priming. Some may argue that by asking participants about cheating so directly we may prime them to associate behavior in the trust game with cheating. To address this concern we ran additional sessions in which, rather than asking our direct cheating notion question above, we asked participants to state how they would feel about various send/return combinations if they were to be assigned the role of sender. The results support the idea that priming is not the driver of reported cheating notions. Details of these sessions and results are provided in Appendix I, section A.2.

Another potential concern with how we elicit cheating notions is that the same individuals who play the game are also asked to report their cheating notions. We made this decision in order to mitigate hypothetical bias stemming from individuals' inability to fully anticipate which outcomes will make them feel cheated without actually playing the game and thereby having pecuniary incentives to understand the consequences of one's own and others' actions. However, some might argue that asking the participants themselves about their cheating notions could bias the reported cheating notions in some other way and that, instead, it would be preferable to ask disinterested parties about what constitutes

[14] In each question "s" and "$f(s)$" were replaced by the appropriate numbers. The words "sender" and "receiver" did not appear on participants' screens.

[15] Our design initially did not include the two explicit opt-out responses mentioned above. Although responding to the question was always completely voluntary, we realized that not providing pre-programmed opt-out responses could make some participants feel obligated to supply a cheating notion even if they did not truly have one. To address this concern, we inserted the two opt-out responses described above. The majority of participants—306 out of 428—took part in sessions featuring the explicit opt-out responses. The remaining 122 participants took part in sessions with no explicit opt-out opportunity. Unless otherwise specified, our analyses utilize all 428 observations. In Appendix I we show that our results are robust to restricting attention only to sessions with explicit opt-out.

cheating. The only study we know of that examines this issue directly in the context of a trust game is Rustichini and Villeval (2012). As part of their study the authors describe a trust game to disinterested parties who then, for two specific send amounts, report the interval of return amounts they would consider fair. These same individuals come back the following week, play the trust game and then again report their fairness intervals. Comparing the lower bounds of these intervals—the closest analogue to the cheating notions we elicit—between disinterested (first week) and involved (second week) parties reveals little difference. Consequently, we feel that having participants report their cheating notions directly after playing the game, while it is still fresh in their minds, is warranted.[16]

2.3 The beliefs elicitation phase

Following the cheating notions questions, participants discovered there would be a beliefs elicitation phase of the experiment and that they could earn additional money according to the accuracy of their estimates. In this phase, each participant was asked to estimate: i) how much other senders would send on average; ii) how much other receivers would return on average; iii) their beliefs about others' beliefs about how much receivers would return (second-order beliefs); iv) other participants' cheating notions; and v) the proportion of other participants who would not cheat them, according to the respondent's own subjective cheating notion (see Appendix II for exact wording).[17] For all belief elicitation questions, participants were instructed to exclude their own actions from their estimates and were told that the accuracy of their estimates would be calculated excluding their own strategies and cheating notions.[18]

[16] A related concern is the lack of financial incentives in our cheating notion elicitation question. Specifically, it has been suggested that we could use the coordination game mechanism of Krupka and Weber (2013), which relies on a clever argument about the focal nature of social norms. While we appreciate their mechanism, our focus on *personal* cheating notions is not amenable to their mechanism. In their own words, they "... provide respondents with incentives not to reveal their own personal preferences but instead to match the responses of others." Moreover, their mechanism is only incentive compatible if two conditions are met: "... a) there is general social agreement that some actions are more or less socially appropriate, constituting the social norm, and if b) respondents attempting to tacitly match others' responses rely on such shared perceptions to help them do so." Neither of these conditions are likely satisfied by personal cheating notions. In the concept of cheating we would like to investigate, individuals should be free to feel cheated irrespective of whether others agree. We realize that the cost of missing financial incentives is, at the very least, added noise, but feel that the benefit of a using a straightforward, transparent, cheating notion question outweighs outweighs this cost.

[17] Items ii)-v) were asked for each possible send amount.

[18] This was done to avoid mechanical correlations between reported beliefs and participants' own strategies or cheating notions.

Participants were informed that one estimate from this section would be chosen to count toward their potential earnings. This chosen belief was remunerated according to a randomized quadratic scoring rule (Schlag and van der Weele, 2013) which is both incentive compatible and theoretically robust to risk preferences. The mechanism was explained to participants in detail. Additionally, participants were told that it was monetarily in their best interest to report their true beliefs and provided with an example illustrating this assertion. An exactly correct belief paid 5 euros in most sessions while, in the remaining sessions, an exactly correct belief paid 20 euros. Beliefs were elicited *after* participants submitted their complete contingent strategies, but *before* knowing their assigned roles.

Eliciting beliefs after game-play and after having elicited cheating notions raises several potential concerns. Central among these are ex-post rationalization of beliefs about others' cheating notions or others' expected returns. For example, participants could ex-post rationalize returning only a little by reporting they believed others expected little back, or by reporting that others needed only a little back in order to not feel cheated. We treat these concerns extensively using several different robustness check exercises. Full details are reported in the Appendix I, Section B.

2.4 Payment phase

After all three phases of the experiment were completed, pairings were randomly determined and within each pair roles were randomly assigned. Outcomes and potential earnings were determined by combining, within each pair, the sender's strategy with the receiver's strategy. Participants were informed at this point which randomly-chosen belief elicitation question would count toward their potential earnings and how much their estimates earned them. Finally, 10% of participant pairs were randomly chosen to be paid their potential earnings.

While 10% may seem low, the experiment was relatively short and convenient, requiring on average about half an hour of participants' time. Furthermore, note that Italian students' opportunity costs are relatively low. As an example, work-study positions at one university in Rome we are familiar with typically pay students around 5 euros per hour. Given both of these observations, we feel the expected earnings from the experiment are commensurate with participants' opportunity cost of time. Despite this, we also conducted a handful of traditional in-lab sessions. We had participants come to the lab and complete the on-line experiment. In these in-lab sessions, 100% of participants were paid their experimental earnings. The patterns in the data from the in-lab sessions were remarkably similar to the patterns in our on-line study data (see Appendix I, Section A.1).

2.5 Instilled values and risk attitudes

For each participant in our main study, we complement the experimental data with data from a previously conducted, unrelated, survey. This survey contains basic demographic information, a (self-reported) measure of the emphasis each participant's parents placed on various normative values during his or her upbringing as well as an incentive-compatible measure of risk aversion (Holt and Laury, 2002).[19]

There was a considerable time lag between the survey and the start of our trust game experiments (from 20 to 60 days) so that survey responses are unlikely to have affected trust game behavior directly. On the other hand, this temporal distance was small enough so that traits, such as risk aversion or instilled values, likely did not change in the meantime. This survey data allows us to control for risk aversion and altruism when examining sender behavior, while instilled values will prove useful in examining what drives receiver behavior.

In Table 1, we summarize key features of the main experiment. Descriptive sample statistics are reported in Table 2.

3 Results

We establish three main results: *i)* there is substantial heterogeneity in cheating notions and beliefs about others' cheating notions; *ii)* intergenerationally transmitted values are important determinants of cheating notions; and *iii)* cheating notions affect decisions on both sides of the trust exchange.

[19] Briefly, this procedure asks participants to make a sequence of ten choices, each of which involves a choice between a relatively risky lottery (38.50 euros with probability p, 1 euro with probability (1-p)) and a safer lottery (20 euros with probability p, 16 euros with probability (1-p)). The probability of the high payoff, p, increases over the sequence from 0.1 to 1.0 in steps of 0.1. This construction implies that more risk averse individuals will switch from preferring the safer lottery to the riskier lottery later in the sequence. We use the choice number in the sequence where this switch occurs as our risk preferences measure which ranges from 1 to 10 and is increasing in risk aversion. For ten percent of survey participants one decision in this sequence was randomly chosen and these participants were paid according to the outcome in their preferred lottery.

3.1 Personal Cheating Notions

We start by remarking that the vast majority of participants—about 80%, averaging across all send amounts—report a personal cheating notion even in sessions where refraining from specifying a cheating notion is salient and simple (see fn 15). Restricting attention to sessions involving explicit cheating notion opt-outs, the proportion of senders selecting the option "this has nothing to do with cheating" ranges from a low of 13 percent when considering sending 10 euros, to a high of 20 percent when considering sending one euro. The proportion of senders who opt out of reporting a cheating notion for *any* reason—which includes selecting either "I don't know,"or "this has nothing to do with cheating,"or just leaving the question blank—in these same sessions is also low, ranging from 17 percent to 23 percent. Apparently, few participants have no opinion one way or the other. Moreover, these patterns suggest that for a large majority of our participants cheating is a well-defined event. These proportions are summarized in Table 3.

Turning from existence to heterogeneity, in Figure 1 we restrict attention to those participants who supply a cheating notion and plot histograms of these cheating notions for several representative send amounts. We overlay each histogram with vertical lines representing two *a priori* plausible cheating notions. The first vertical line represents the cheating notion most commonly assumed in the trust literature: a weakly positive return on investment (WPROI) rule.[20] An individual defining being cheated according to this rule would for each send amount, $s \in \{1, \dots, 10\}$, report a cheating threshold of exactly s, feeling cheated for any return amount strictly less than s but not feeling cheated for any return amount weakly greater than s. The second vertical line represents an equal split of the receivers' entire earnings—i.e., for each s, the line is placed at $\frac{f(s)}{2}$. Accordingly, we call this an "equal split" (ES) cheating notion. However, recall that since our design features unequal initial endowments, this notion should not be confused with inequality aversion or egalitarianism. Instead, demanding half of the receivers earnings typically implements a lot of inequality in final earnings.[21] One justification for an ES cheating notion is that individuals may generally feel entitled to an equal share of all of the money their actions

[20] This is the cheating standard explictly assumed in Berg, Dickhaut and McCabe, 1995 and incorporated into much of the subsequent literature on trust (*cf.* Bohnet and Zeckhauser, 2004).

[21] For example, if the sender sends 1 euro, the receiver receives 8.05 euros. An individual with an ES cheating notion would feel cheated by receiving less than 4.02 euros back which would correspond to (sender earnings, receiver earnings)$= (14.02, 4.03)$.

generate, which *could* be interpreted as the receiver's entire earnings.

As is evident from the histograms, there is quite a lot of heterogeneity in personal cheating notions, suggesting that the typical ad-hoc assumption of a uniform standard of cheating in trust-based exchange is unwarranted. The histograms suggest that cheating notions are, instead, roughly bimodal with much of the mass concentrated between WPROI and the typically much more demanding ES cheating notion. Consequently, while WPROI may serve well as a lower bound on behavior generating the feeling of being cheated, a lot of information on individual heterogeneity is lost by assuming that it *is* most people's cheating notion.

To get a more quantitative feel for the heterogeneity in cheating notions engendered by trust-based exchange, we next classify participants according to whether their personal cheating notions are consistent with a return on investment rule or an ES rule. We first consider each send amount, $s \in \{1, \ldots, 10\}$, separately and then go on to ask whether individuals are consistent with their definitions across send amounts. For comparison with previous literature, we also consider whether cheating notions are consistent with an egalitarian rule (inequality aversion) according to which cheating would be defined with respect to whether final money outcomes are equal. As before, we restrict attention to participants supplying a cheating notion who, in any event, represent the bulk of our participants.

In the first row of Table 4 we report the proportion of participants specifying WPROI as their personal cheating notion.[22] As expected, we find that WPROI performs poorly, describing only a small fraction—from 10% to 18%—of our participants' cheating notions. From 70% to 80% of our participants report cheating notions *strictly* larger than WPROI.

To be a bit more generous to the idea that participants define cheating according to return on investment, we next construct a related measure allowing for a reasonable ($r = 10\%$) *strictly positive* return on investment (SPROI), taking into account whether or not there was a 0.50 sending fee. To accomodate experimental participants' well-known proclivity for responding with whole numbers, we next calculate the least integer greater than this value—say, n—and classify as consistent with SPROI any cheating notion falling within the closed interval $[s, n]$.[23] This more lenient return on investment notion fares slightly

[22] Here, to be generous to this cheating notion, in sessions with a sending fee of 0.50, WPROI includes both participants who report s and participants who report $s + 0.50$ as their cheating threshold

[23] As an example, consider a session with a 0.50 sending fee. For $s = 5$, we would first calculate $(s + 0.50) \times (1 + r) = (5.50) \times (1.1) = 6.05$. The least integer greater than 6.05 is 7. Therefore, any cheating notion falling in the interval $[5, 7]$ would be labeled as consistent with SPROI.

better, accounting for anywhere from 18% to 36% of cheating notions. Still, the majority of subjects who report a cheating notion—from 50% to 60%, depending on s—report one that is strictly higher than even SPROI suggesting that a lot is lost by assuming that individuals uniformly define cheating according to a reasonable return on investment.

In the next row of Table 4, we report the proportion of participants whose cheating notions are consistent with one popular model of social preferences: inequality aversion (Fehr and Schmidt, 1999). For each send amount separately, we label as being consistent with inequality aversion any cheating notion falling within the smallest interval with integer endpoints surrounding an exactly equal division of total available surplus.[24] As is evident from the reported proportions, literal inequality aversion consistently characterizes very few of our participants' cheating notions.

Finally, in the last row of Table 4 we report the proportion of participants who apparently demand an equal share of the fruits of their action—i.e., half of the receivers' entire earnings—in order to not feel cheated (ES). We again take into account participants' proclivity for integer-valued answers by labeling any cheating notion that falls within the closed interval bounded by the largest integer less than, and the smallest integer greater than, an exactly equal split of the entire amount receivers receive.[25] We find that ES consistently accounts for around one-third of our participants' cheating notions.

Summarizing our findings so far, we find that there is a lot of heterogeneity in personal cheating notions and that their distribution is roughly bimodal. The two most prevalent cheating notions call for either a strictly positive return on investment—SPROI—or fully half of the fruits of one's actions—ES. What we have not seen from the data is whether specific individuals typically define cheating consistently across possible send amounts. Such stability would provide reassurance that reported cheating notions reflect some underlying individually stable trait.

Toward this end, we first restrict attention to individuals whose cheating notions were classified as consistent with an ES cheating notion for a send amount of 1—the send amount

[24] For example, consider a high fee session with a sending fee of 0.50. Further, consider $s = 2$. The total surplus available in this case is $10.50 - 0.50 - 2 + 11.30 = 19.30$, and half of this surplus is 9.65. We would label as consistent with inequality aversion any cheating notion in the interval $[9, 10]$. This assumes, of course, that cheating notions reflect only the disutility experienced from inequality and not the standard, pecuniary, portion of utility.

[25] To be clear, consider the send amount $s = 3$. If the sender sends 3, the receiver receives $f(s) = 11.30$. An exactly equal split of 11.30 entails the receiver returning $\frac{f(s)}{2} = 5.65$. Consequently, for $s = 3$ we would label as an equal-split cheating notion any cheating notion falling within the interval $[5, 6]$.

providing the widest separation between ES and SPROI. We then plot histograms of these individuals' cheating notions for several other representative send amounts from the set $s \in \{2, \ldots, 10\}$. We repeat this exercise for individuals who were classified as SPROI for a send amount of 1. We superimpose on each histogram two vertical lines: one at the send amount, s, and one at exactly half of the receivers' earnings ($\frac{f(s)}{2}$). These histograms are reported in Figure 2 where we see a remarkable amount of consistency for individuals defining cheating according to ES (top panel). We find somewhat less, but still substantial, internal consistency for individuals who define cheating according to return on investment definition (bottom panel).

Having documented both the heterogeneity and individual consistency of the cheating notions engendered by trust-based exchange, we next ask whether these features carry over to beliefs about others' cheating notions. With this goal in mind, we plot three sets of histograms. The first set of histograms are simple plots of participants' beliefs about others' cheating notions (Figure 3). As with own cheating notions, we overlay each histogram with vertical lines placed at s and $\frac{f(s)}{2}$, corresponding to WPROI and ES cheating notions, respectively. We see that beliefs follow much the same distribution as own cheating notions, being characterized by a lot of heterogeneity with mass concentrated between WPROI on the lower end and ES on the most demanding end.

In Figure 4, we plot two more sets of histograms. In the top panel, we restrict attention to participants whose own cheating notions classified them as ES for send amount 1 and plot histograms of these participants' beliefs about others' cheating notions for a representative subset of send amounts, $s \in \{2, \ldots, 10\}$. In the bottom panel we repeat this exercise restricting attention to participants whose own cheating notions were consistent with SPROI when sending 1 euro. For both sets, we overlay each histogram with vertical lines at WPROI and ES. From this figure we can glean two features of cheating notion beliefs. First of all, as with cheating notions themselves, there is considerable consistency of cheating notion beliefs across send amounts. Secondly, individuals tend to expect a positive relationship between their own cheating notions and others' cheating notions: individuals who define cheating according to ES have beliefs about others' cheating notions concentrated around ES; individuals who define cheating according to a SPROI have beliefs concentrated around a return on investment notion of cheating. This pattern is consistent with a false consensus effect (Ross, Green and House, 1977): individuals tend to believe others are similar to themselves.

3.2 What Determines Cheating Notions?

So far our data suggest the existence of a substantial heterogeneity in cheating notions and beliefs about others' cheating notions. We also documented that individuals tend to expect a positive relationship between their own cheating notions and others' cheating notion. We try to understand the reason for this link, emphasizing the relevance of a substantial stable, culturally transmitted component.

We start with a plausible conjecture based on previous research about belief formation: in novel situations introspection substitutes for information so that through the well-established psychological phenomenon known as "false consensus" one's own cheating notions become a significant determinant of beliefs about others' cheating notions (see, e.g., Ross, Green and House, 1977; in a trust game context, see Butler, Giuliano and Guiso, 2012). If own cheating notions themselves are persistent—perhaps being based on moral values which tend to be culturally transmitted from parents to children (see e.g. Bisin and Verdier, 2010)—then cheating notion beliefs may also persist over time and context. There are two links in this chain: i) cheating notion beliefs to cheating notions; ii) cheating notions to values. We provide evidence on both links.

On the first link, there is an abundance of evidence in our data suggesting that own cheating notions contribute substantially to cheating notion beliefs. For example, the raw correlation between own cheating notions and cheating notion beliefs is both large in magnitude and highly statistically significant for all send amounts. The correlation ranges from a low of 0.55 to a high of 0.66 and is always significant at greater than a 1% level.[26] Moreover, regressing cheating notion beliefs on own cheating notions for each send amount separately, while controlling for available demographics, paints a similar picture. In these regressions (omitted, but available on request) the coefficient on own cheating notions ranges from 0.52 to 0.57 and is always significant at better than a 1% level.

To investigate the second link, we test directly for a relationship between our parti-

[26] With regard to own cheating notions, we run into a technical problem: how to handle participants who refrain from supplying a cheating notion. One potential answer is to code the cheating threshold as 0 whenever a participant responds "this has nothing to do with cheating," and as missing if they fail to supply a cheating notion for any other reason. This is what we do. An alternate strategy of coding the cheating notion as missing whenever participants fail to report a cheating threshold for *any* reason yields similar conclusions.

cipants' cheating notions and the values their parents emphasized during their upbringing while controlling for our standard set of demographic variables. We use data from a previously conducted unrelated survey (described in Section 2.5) which included a section on parentally instilled values. The survey asked about a rich set of normative values. For each normative value in this set, survey participants were asked to state how much emphasis their parents placed on this value during their upbringing which we take to be a proxy for received cultural values.[27] Valid responses ranged from 0, which indicates no emphasis, to 10 which indicates quite a lot of emphasis.[28] For our estimates, we select a relevant subset of these normative values and organize them into two categories: "cooperative" and "competitive." The former category includes such values as helping others and honesty. The latter category includes, for example, the value of striving to be better than others.[29] We construct an index of parents' emphasis on "cooperative" and "competitive" values by taking the average emphasis over all the values constituting each category. This yields a measure for each category theoretically ranging from 0 to 10. We divide each of these measures by 10 obtaining an index on a 0 to 1 scale.

To get a summary measure of the relationship between instilled values and own cheating notions, we pool over all send amounts and regress cheating notions on cooperative and competitive values. Since pooling in this manner results in multiple observations for each participant we incorporate individual-level random effects in our model. As the presence of an investment fee may directly affect cheating thresholds we also include a dummy for sessions with no investment fee. Finally, because our trust production function is approximately quadratic in money sent, we allow cheating notions to be a quadratic function of money sent.

The estimates reported in Table 5 reveal a substantial relationship between values and cheating notions. Interestingly, our data suggest that the two classes of values we consider

[27] We acknowledge that such self-reported retrospective questions are likely to be noisy or biased measures of the values our participants' parents *actually* emphasized. For example, individuals may selectively remember some lessons and not others, biasing their recollection of what their parents taught them. Unfortunately, our data do not allow us to address this criticism directly since we do not survey our participants' parents. However, it is reasonable to assume such self-reports convey some information about the values our participants *believe* their parents transmitted to them, which should lend some credence to our interpretation of them as *received* cultural values.

[28] Participants could also respond "I don't know," which we code as missing.

[29] The full set of "cooperative" values is: i) behave as a model citizen; ii) help others; iii) group loyalty; iv) always give others their fair share; v) always tell the truth; vi) always keep your word. "Competitive" values are: i) always extract the maximum advantage from every situation; ii) seek to be better than others; iii) act so as to induce good in others (e.g., scold somebody who litters).

pull in opposite directions. Instilled cooperative values significantly lower cheating notions: the more emphasis parents placed on cooperative values, the fewer euros senders need back in order to not feel cheated. Competitive values, on the other hand, have the opposite effect, raising cheating notions significantly. Controlling for instilled values, cheating notions tend to move one-for-one with the amount sent, suggesting that a positive return on investment rule is the baseline cheating notion and that values determine how far individuals deviate from this baseline. Finally, there is little evidence that cheating notions vary by demographics once we control for values.

Summing up, our data suggest that parentally instilled values are significant predictors of cheating notions and that cheating notions, in turn, are highly significant predictors of cheating notion beliefs, lending some credence to the idea that cheating notions and related beliefs are stable predictors of behavior. Consequently, for the remainder of our study we focus on the relationship between cheating notion beliefs and behavior.

3.3 The Relationship between Cheating Notion Beliefs and Behavior

In this section we look at the effect of cheating notion beliefs on the behavior of both receivers and senders.

3.3.1 Receivers' Decision to Intentionally Cheat

One advantage of focusing on cheating notion beliefs directly is that we can study what drives receivers' decision to intentionally cheat. We can address this latter question directly because we know when receivers cheat according to their own estimates of others' cheating definitions.

We construct a dummy variable taking the value of 1 whenever receivers return less than they themselves believe their co-players need back in order to not feel cheated and 0 otherwise, for each amount a sender could send. This dummy indicates when receivers intentionally cheat. We then relate this variable to receivers' demographic characteristics as well as their own, and their estimates of others', cheating notions.[30]

[30] Cheating notions for participants who refrain from supplying a cheating notion were coded as 0 whenever they selected "this has nothing to do with cheating," and as missing if they failed to supply a cheating notion for any other reason. An alternate strategy of coding the cheating notion as missing whenever participants fail to report a cheating threshold for any reason roughly doubles the magnitude and increases the significance of all own cheating notion coefficients. The same happens simply including a dummy for those who report that sending s euros has “nothing to do with cheating.” The results presented should therefore be seen as a conservative estimate of the impact of own cheating notions.

Table 6 presents our estimates of receivers' propensities to intentionally cheat for each possible send amount. Participants' demographics have few consistent effects on cheating across different send amounts: older participants generally cheat less for lower send amounts; smarter participants—those who have higher math scores—are less likely to cheat for high send amounts. Interestingly, gender plays no role. On the other hand, controlling for receivers' expectations about senders' cheating notions, receivers that have higher standards—i.e., who would feel cheated unless they were given back a lot when playing as senders—are consistently less likely to cheat across all send amounts. We interpret this finding as saying that more demanding people tend to refrain from cheating others, behaving according to the principle "do not do to others what you would not want others to do to you." Notice, however, that conforming to this principle is cheaper when amounts sent are low and the temptation to deviate from it (and doing to others what you would not want them do to you) is thus weaker. Consistent with this we find that the effect of one's own cheating notion is stronger at low levels of amount sent and weaker at high levels: the reported probit coefficients imply that the marginal effect of an increase in receivers' own notions of cheating at send amount 10 is half that at send amount 1 (1.6 percentage points vs 3.6 percentage points, respectively).

3.3.2 The Effects of Cheating Beliefs on Senders' Behavioral Trust

As a second step, we consider whether and how the specter of being cheated affects senders' behavior. While previous research suggests that expected cheating or betrayal may affect the likelihood of trusting behavior (e.g., Bohnet and Zeckhauser, 2004), it is an open question whether cheating is an important determinant of the intensive margin of trust—i.e., *how much* to trust, conditional on trusting at all. This is an important distinction as it speaks to the potential benefits that may obtain in terms of surplus creation from policies aimed at reducing cheating. For example, if expected cheating determines the extensive margin of trust only, then there may be little to gain from reducing cheating in already highly trusting environments.

To examine whether expected cheating affects the intensive margin of trust, for each participant we construct a unidimensional measure of his or her beliefs about the proportion of non-cheaters in the (experimental) population. We do this by averaging each participant's answers to the following set of 10 questions ($S = 1, 2, \ldots, 10$):

> "If you send s euros and B therefore receives $f(s)$ euros, what percent of B's will return enough money so that you do not feel cheated?"

The resulting measure of beliefs about population trustworthiness theoretically ranges from 0 to 1, with 1 indicating the sender believes no receiver will cheat for any send amount (all are trustworthy) and 0 indicating all receivers will cheat for every send amount (none is trustworthy).

Before proceeding we must address one technical issue: how to construct this measure for individuals who, for a particular amount sent, report no cheating notion. First of all, if an individual responds that sending S euros "...has nothing to do with cheating," then it is reasonable to assume that this individual *cannot* feel cheated regardless of the receiver's decision. Therefore, we code such individuals' population trustworthiness belief conditional on sending s euros as 1 before constructing the summary measure.[31] On the other hand, if an individual did not report a cheating notion conditional on sending s euros for any other reason, then our elicitation mechanism is not incentive compatible since we cannot observe whether such an individual will feel cheated. For these participants, we code as missing their answer to the population trustworthiness question associated with sending s euros, which also results in a missing observation with respect to our summary measure.

Given these caveats, we construct a unidimensional measure of beliefs about population trustworthiness for 401 (out of 428) participants, which we interpret as their subjective probabilities of not being cheated. Figure 5 plots the kernel density of this probability separately for opt-out and no-opt-out sessions. We document a modal value at around 0.5 (almost equal to the fraction of non cheaters in the pool—see Table 2, bottom row) irrespective of opt-out opportunities. In sessions with opt-out (the dashed line), a second mass of observations centers around a value of 1, reflecting (mechanically) the small minority of participants who report the trust game "has nothing to do with cheating" consistently.

In an analogous fashion, we construct for each participant a summary measure of his or her beliefs about the proportion of the money they send that will be returned to them. For each $s \in \{1, \ldots 10\}$ we divide the participant's estimated return *amount* conditional

[31] This could be problematic for our analysis if the subset of people who consistently report that sending money has nothing to do with being cheated also sends more on average. However, this does not seem to be the case. Only 39 individuals have a population trustworthiness measure equal to 1. The average send amount for these 39 individuals is 4.28, which is not significantly different from the average send amount for the remaining 362 individuals (4.34).

on sending s euros by s to get their estimated (gross) return proportion. We then average their 10 return proportion estimates to get a unidimensional measure of return proportion beliefs. The resulting averages range from a low of 0.00 to a high of 4.02 with a mean of 1.27 and a standard deviation of 0.64. We interpret this index as a measure of senders' expected (gross) return proportion and note that, on average, expectations are nearly identical to actual gross return proportions (Table 2).

Finally, using these two summary measures we estimate a model of how much senders send as a function of the senders' expected return proportion, their beliefs about being cheated and an interaction between these two variables. We control for our standard set of demographics. To account for selection into sending a positive amount we estimate a Heckman model and exploit variation in the investment fee across sessions to construct the selection equation. Specifically, the exclusion restriction for the selection equation consists of a dummy for "Low fee" sessions where the investment fee was zero. Importantly, because two common alternative explanations for senders' behavior in the trust game are risk preferences and altruism, among our demographic controls we include an incentive-compatible measure of risk aversion collected from the survey described in Section 2.5 as well as a proxy for altruism obtained from that same survey.[32]

Table 7 presents the estimates. The second column presents the selection equation, which is a probit model estimate of the decision to send something versus nothing (i.e., the extensive trust margin). As desired, this extensive margin depends significantly on the presence of a sending fee. The first column presents the main equation which estimates the intensive margin of trust formally accounting for selection into sending a positive amount. Here, the estimate implies that the specter of being cheated plays a significant role in the intensive margin of trust: the positive and significant coefficient on our measure of the expected probability of *not* being cheated indicates that when senders believe it is less likely that they will be cheated, they send more. The implied effect of non-cheating beliefs on behavioral trust is large: increasing this belief from 0.1 to 0.9 is associated with an increase in the average amount sent equal to 51% of the sample mean. The coefficient on expected pecuniary returns is also positive and significant, indicating that standard pecuniary concerns also drive senders' behavior. Finally, the negative and (marginally) significant coefficient on

[32] We use as our measure of altruism the emphasis, on a scale from 0 to 10, participants' parents placed on the value of "helping others."during their upbringing.

the interaction between expected returns and non-cheating beliefs suggests that as expected pecuniary returns increase, the negative impact on trust of expected cheating subsides. In other words, the sting of expected betrayal can be soothed by money.[33]

4 Discussion and interpretation

In this section we try to put our results in perspective. We first link our results to theoretical models of guilt aversion. We then provide a more general view on the type of preferences that could explain receivers' cheating decision. We focus on the essential points here but further details are provided in the appendix (part D).

4.1 Cheating and guilt aversion models

A central piece of guilt aversion theory (Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007) is the relevance of first and second order beliefs. In this literature, guilt is the result of disappointing others with respect to their formal, mathematical, expectations of counter-party behavior: person A is disappointed whenever person B's actions fall short of A's expectations of B's actions. Consequently, B's second-order beliefs—B's beliefs about A's beliefs—shape the set of possible equilibria. In this section, we investigate whether cheating notions are an important source of expectations about how others will behave. If first-order (second-order) beliefs about others' actions (beliefs) are closely empirically related with personal cheating notions, then knowledge about the distribution of personal cheating notions in a population can provide insight into which of the multiple equilibria typically predicted by guilt theoretical models are most likely to occur.

Intuitively second order beliefs and beliefs about others' cheating notions may come to be connected through several channels. For example, as most of our daily interactions do not involve being cheated, induction or Bayesian updating may lead individuals to formally *expect* to not be cheated.[34] Alternatively, correlation between senders' cheating notions and

[33] The results are virtually the same if we estimate a Tobit model of send amounts, which intuitively models selection as censoring.

[34] We provide evidence in Appendix I, Section C, that reported cheating notions are not "reverse-caused" in this respect: i.e., that participants do not form beliefs about the amounts participants return and then simply report this belief as their cheating notion so as to avoid, e.g., looking liking a sucker. Essentially, we show that cheating notions are no more correlated with beliefs for outcomes which may actually happen—where looking foolish is a possibility—than for outcomes that are impossible.

their beliefs about receivers' actions can be generated from a simple fixed cost of cheating model with shared knowledge of cheating notions. If, for whatever reason, senders' beliefs about receivers' actions correlate with senders' cheating notions, then receivers' second-order beliefs about senders' beliefs should rationally reflect this correlation.[35]

We test for the conjectured correlations between i) own cheating notions and beliefs about receivers' actions; and ii) beliefs about others' cheating notions and receivers' second order beliefs[36]. The main lesson from our exercise is that one's own cheating notion is consistently a highly significant predictor of senders' first-order beliefs and that beliefs about others' cheating notions exhibit a strong positive relationship with second-order beliefs.

Having seen that receivers' cheating notion beliefs and their second-order beliefs are closely related empirically, the question arises: do second-order beliefs contain predictive power for receivers' behavior beyond what is contained in cheating notion beliefs? We find that, controlling for cheating notion beliefs, the relationship between receivers' behavior and their second-order beliefs is typically non-significant. On the other hand, beliefs about others' cheating notions are always significant predictors of receivers' behavior.[37]

Using the definition of intentional cheating (defined in section 3.3.1), we can also test a basic prediction of how guilt arising from cheating should constrain receivers' behavior: those who refrain from intentionally cheating should do so minimally, as returning more than necessary to avoid cheating does not reduce guilt but does reduce own money earning. To do so, we split the sample between cheaters and non-cheaters and estimate the amount receivers return as a function of their beliefs about senders' cheating notions and our standard set of demographics. To formally account for selection into cheating or not cheating, we estimate Heckman models using as their exclusion restrictions in the selection equations participants' own cheating notions. The results (Table 8) are broadly consistent with cheating giving rise to guilt.[38] For those who choose to refrain from cheating, return amounts

[35] It is worth noting that senders' beliefs about receivers' actions and senders' cheating notions are clearly conceptually distinct notions: the latter are value judgments about particular behaviors, while the former are mathematical assessments about the probability of particular events. In principle, there is no reason one cannot judge normatively bad outcomes (being cheated) to be likely outcomes, or unlikely outcomes to be good. Theoretically, then, there is little reason to predict that cheating notions and beliefs about receivers' actions will be correlated. By similar logic, receivers' beliefs about senders' beliefs about receivers' behavior—receivers' second order beliefs—should be conceptually distinct from, and not necessarily correlated with, receivers' beliefs about senders' cheating notions.

[36] For details about the empirical strategy to test for these correlations and the corresponding results, see part D of the appendix and Tables A14-A15.

[37] See Table A16 of the appendix.

[38] Ignoring selection issues and estimating simple OLS models of return amounts yields qualitatively similar

vary essentially one-to-one with their beliefs about cheating notions, suggesting that for non-cheaters, cheating notion beliefs are acting as thresholds. On the other hand, receivers who cheat their co-players are much less sensitive to cheating notion beliefs. The estimated coefficients on these beliefs are consistently around half as large as for non-cheaters.

4.2 A more general view on receivers' cheating decisions

In this section we try to provide a more general view on the type of preferences that could explain receivers' cheating decision.

We start by plotting (Figure 6) the fraction of receivers who cheat at each send amount after partialling out the effect of the expected notion of cheating, thus purging the data from the mechanical effect this has on the probability of cheating. The share is 38% at send amount 1 and decreases, roughly, monotonically down to 1% at send amount 10.

This declining propensity to cheat as receivers receive larger sums from senders is inconsistent with both purely selfish preferences and fixed-cost of cheating models. With purely selfish preferences receivers would always cheat. On the other hand, since potential pecuniary gains from cheating increase in the amount sent, fixed cost of cheating models would predict a *non-decreasing* relationship between amount sent and cheating propensity.[39]

Patterns in our data also appear to be inconsistent with literal interpretations of many influential social preferences models. For example, consider inequality aversion (Fehr and Schmidt, 1999) or social welfare preferences (Charness and Rabin, 2002). Inequality averse individuals lose utility from unequal outcomes, while individuals with social welfare preferences place weight in their utility calculations on the outcome of the worst-off individual in their reference group as well as the total amount of money being distributed. In a two-player decision-making setting with no surplus creation opportunity such as that faced by trust game receivers, both of these models predict that receivers should never willingly put themselves behind in terms of final monetary payoffs.[40] However, a large fraction of receivers in our study do exactly that. For example, 82% of receivers willingly put themselves

results.

[39] Let $B(S)$ denote the pecuniary benefit to the receiver from cheating, which increases with S, and assume the receiver j's fixed cost of cheating, K_j, is randomly drawn from the distribution $F(K)$. Then as S (and $B(S)$) increases, there should be a (weakly) higher proportion of receivers for which $B(S)$ exceeds K_j, and who therefore cheat.

[40] It is, of course, true that this strong prediction fails to hold if the receiver's reference group is something other than *just* himself or herself and his or her co-player. What the reference group is, or should be, is an important open question outside of the scope of this paper. We follow most of the literature in assuming participants view the trust game as a two-person interaction.

further behind than necessary when sent 1 euro and 47% of the receivers put themselves behind when sent 4 euros.[41]

The estimation results in Table 8 and the cheating pattern in Figure 6 could be justified in (at least) two ways. The standard justification is positive or negative reciprocity: sending more is a nicer action and/or sending less is a meaner action, so reciprocity demands responding in kind with a nice action (not cheating) or a mean action (cheating). An alternative explanation comes directly from the definition of trust: trust entails *vulnerability*. At the same time, a widespread and intuitive moral standard is that, irrespective of what constitutes cheating, it is *more* wrong to cheat the more vulnerable. For example, cheating the elderly or the very young is commonly viewed as particularly reprehensible. This is the point made by Gneezy (2005). In the context of the trust game, sending more makes senders more vulnerable. Consequently, it is reasonable to assume that the moral costs of cheating increase in amount sent.

Without additional information we can rule out neither reciprocity nor vulnerability as the driving motive. To shed some light on why cheating declines in amount sent, in one of our robustness treatments we asked participants to describe their rationale for how they played the role of receiver in the trust game (Appendix I, section A.2.1). They could select one response from among four pre-programmed options and one free description. Three of the pre-programmed options were meant to capture positive reciprocity, negative reciprocity and vulnerability motives, respectively, while the fourth was essentially a "decline to state." option. We found that the modal response—selected by 42 percent of participants (72 out of 170) was the vulnerability explanation. The second most common response was positive reciprocity (about 30 percent of participants), while almost nobody chose negative reciprocity (6 percent; 10 out of 170 participants).

In light of these patterns, a unified way to model both senders' and receivers' preferences that is consistent with our data is to augment standard pecuniary preferences with a moral cost function. Individuals incur disutility from immoral actions, either when they are the perpetrator or the victim of such actions. Receivers lose utility when they cheat. Senders lose utility when receivers cheat them, which is consistent with our finding that expected

[41] This ignores the extra 50 cents senders have in sessions with no investment fee. Taking into account this extra 50 cents could obviously only increase these percentages. It should also be noted that receivers need not put themselves behind, except when the amount sent is 1. But even there, the figure of 82% includes only those receivers returning a *strictly* positive amount, so these receivers are willingly putting themselves further behind their counterpart than necessary.

cheating has a direct negative impact on the amount senders send. Beyond implying disutility from being cheated, our data do not say much about what senders' moral cost function might look like. On the receiver side, however, our data provide a bit more bite. For the rest of this section, therefore, we focus on receivers' preferences.

As a flexible specification for receivers' moral cost function, we assume it has three arguments: the vulnerability of the sender as measured by the amount sent, s; a fixed cost of cheating term; and a term measuring the degree with which the receiver cheats, as defined by the distance between the receiver's estimate of the sender's cheating notion and the amount the receiver returns, r.

In general, denote this moral cost function $m(s, K_j, dist(r, c_j(s)))$. A receiver's utility is then given by:

$$U_j(r, s, c_j, K_j) = u(f(s) - r)) - \mathbb{I}(r < c_j(s)) \times m(s, K_j, dist(r, c_j(s))) \tag{1}$$

In (1), $f(s)$ denotes how much the receiver receives when the sender sends s and $\mathbb{I}(r < c_j(s))$ is an indicator function taking the value of 1 whenever the receiver intentionally cheats by returning less than dictated by the receiver's own estimate of the sender's cheating notion, $c_j(s)$. We assume that m is increasing in s, the vulnerability of the sender. We also assume that the fixed cost of cheating, $K_j \geq 0$, is a random draw from a common non-degenerate distribution function, $F(K)$. Finally, we assume that m is increasing and convex in its last argument, $dist(r, c_j(s))$, so that higher degrees of cheating are increasingly morally costly.

To be more concrete, a simple utility specification satisfying these assumptions is given by:

$$U_j(r, s, c_j, K_j) = u(f(s) - r)) - \alpha_j \mathbb{I}(r < c_j(s)) \times \{K_j + v(s) + \gamma_j (c_j(s) - r)^2\} \tag{2}$$

In equation 2, we assume that $u(f(s) - r)$, the receiver's standard pecuniary utility, is increasing and concave. The rest of the utility function captures the moral cost of cheating. The parameter α_j captures how much receiver j cares about morality. The parameter γ_j captures how much the receiver cares about degrees of cheating. A sender's vulnerability or niceness is captured by $v(s)$ which we assume is increasing.

There are three points to notice about this utility specification. First of all, setting $\alpha_j = 0$ reduces receivers' utility to standard (amoral) preferences; Secondly, notice that whenever $\alpha_j > 0$, setting $\gamma_j = v(s) \equiv 0$ implies receivers have simple fixed-cost-of-cheating preferences. Finally, if we assume that receivers expect senders to expect not to be cheated, then receivers' beliefs about senders' cheating notions—$c_j(s)$ in our model—may be closely linked to receivers' (second order) beliefs about how much money senders' expect receivers to return. In this case, our model can be thought of as an alternative way to capture guilt aversion which does not require knowledge of second order beliefs. Supporting this view, as already noted, in our data receivers' second-order beliefs are highly significantly correlated with their beliefs about others' cheating notions.

The specification for receiver utility given in equation 2 can explain: a) why the decision to cheat depends on others' expected cheating notions; b) why cheating depends on the intensity of moral preferences as proxied for by receivers' own cheating notions; and c) why the probability of cheating decreases in amounts sent as shown in Figure 6. This latter feature would be implied, for instance, whenever there are sufficiently many receivers with $\alpha_j > 0$ and when $v(s)$ is sufficiently steep in s. Intuitively, as $v(s)$ becomes steeper, cheating more vulnerable senders requires a larger offsetting pecuniary utility gain.

This simple preference specification can also account for another feature of the data: conditional on cheating, receivers on average do not go so far as to return nothing. Instead, they send *something* back. In our model, the amount returned by cheaters should depend positively on expected cheating notions, but—and this is the key prediction—it should *not* move one-to-one with the expected notion of cheating. On the other hand, conditional on not cheating, receivers should return the minimum amount consistent with satisfying the sender's notion. Non-cheaters' return amounts should therefore move one-to-one with the expected cheating notion.[42] Only the latter prediction is shared by both our model and the fixed cost of cheating model. As we have seen, both predictions find support in our data.

5 Concluding Remarks

Many real life exchanges require the "trustor" to decide whether and how much to trust a "trustee" who makes no promise on how he will behave in response to the trust received. This paper investigates what individuals' personal, subjective notions of what constitutes cheating can tell us about behavior in such situations. Our study takes place in the context of a trust game where we elicit participants' definitions of being cheated and a wide array of related beliefs.

In this context, our data suggest several patterns. First of all, participants have personal cheating definitions when playing the trust game. We find that these personal cheating notions and beliefs about others' cheating notions are quite heterogeneous but roughly bimodal, clustering around an equal-split rule and a positive return on investment rule. We also find evidence consistent with cheating notions being culturally transmitted, and hence stable, which is important since we also find that stability in own cheating notions may

[42] The receiver's optimal choice can be found as follows. Suppose the receiver decides to cheat so that his or her utility is

$$U_j(r, s, c_j, K_j) = u(f(s) - r) - \alpha_j \times \{K_j + v(s) + \gamma_j (c_j(s) - r)^2\} \tag{3}$$

The amount the receiver sends back, r^*, is given by:

$$u'(f(s) - r^*) = 2\alpha_j \gamma_j (c_j(s) - r^*) \tag{4}$$

and is increasing in the estimated cheating notion $c_j(s)$ with a slope that is less than 1.
If the receiver decides not to cheat, the utility obtained is

$$\begin{aligned} & u(f(s) - r) && (5) \\ \textit{subject to}: \; r \quad \geqq \quad & c_j(s) && (6) \end{aligned}$$

and is maximized by setting $r = c_j(s)$ so that when the receiver does not cheat, the amount returned varies one-to-one with the expected cheating notion.
Finally, the receiver decides whether or not to cheat by comparing utility under the two cases and thus cheats if

$$u(f(s) - r^*(c_j(s))) - \alpha_j \times \{K_j + v(s) + \gamma_j (c_j(s) - r^*(c_j(s))^2\} > u(f(s) - c_j(s)) \tag{7}$$

or

$$u(f(s) - r^*(c_j(s))) - u(f(s) - c_j(s)) > \alpha_j \times \{K_j + v(s) + \gamma_j (c_j(s) - r^*(c_j(s))^2\} \tag{8}$$

where the left hand side is the net utility gain from cheating and the right hand side is the moral cost of cheating. This expression makes it clear that as s increases, provided $v(s)$ is sufficiently steep, cheating will diminish.

translate into cheating notion belief stability through false consensus. Finally, we provide evidence that cheating notion beliefs affect the behavior of both sides to the exchange. All together, our results suggest that studying cheating notions and related beliefs can help us understand and predict behavior in trust-based exchange.

An interesting question which we cannot address with our current data is how *knowing* that there are multiple notions of cheating affects sender and receiver behavior, either in the one-shot context here or when, more realistically, individuals interact repeatedly. One may wonder whether individuals adapt their own cheating notions to be more in line with the average population cheating notions causing an eventual convergence to one normative cheating standard; or, rather, whether those with high cheating notions cease to interact with the general population because they feel cheated more often in their interactions. We leave these and related questions for future research.

References

[1] Akerlof, George A. and William T. Dickens (1982), ”The Economic Consequences of Cognitive Dissonance,” *The American Economic Review*, 72, 307-319.

[2] Balafoutas, L., A. Beck, R. Kerschbamer, and M. Sutter (*forthcoming*), “What drives taxi drivers? A field experiment on fraud in a market for credence goods.” *Review of Economic Studies*.

[3] Battigalli, Pierpaolo and Martin Dufwenberg (2007). “Guilt in Games,” *American Economic Review*, 97, pp. 170-176.

[4] Berg, J., Dickhaut, J. and K. McCabe (1995), ”Trust, Reciprocity and Social History,” *Games and Economic Behavior*, 10, 122-142.

[5] Bohnet, Iris and Richard Zeckhauser (2004), “Trust, Risk and Betrayal.” *Journal of Economic Behavior and Organization*, 55(4), pp. 467-484.

[6] Bisin, Alberto and Thierry Verdier (2010), “The Economics of Cultural Transmission and Socialization.” In Jess Benhabib, Alberto Bisin and Matthew O. Jackson editors: Handbook of Social Economics, Vol. 1A, The Netherlands: North-Holland, 2011, pp. 339-416.

[7] Butler, Jeffrey V., Paola Giuliano and Luigi Guiso (2012), “Trust, Values and False Consensus.” NBER Working Paper No. 18460

[8] Castillo, Marco, Ragan Petrie, Torero Ragan, Maximo A. Torero and Lise Vesterlund (2012), “Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination.” NBER Working Paper No. w18093

[9] Charness, Gary and Martin Dufwenberg (2006). “Promises and Partnership.” *Econometrica*, 74(6), pp. 1579-1601.

[10] Charness, Gary and Matthew Rabin (2002). “Understanding Social Preferences with Simple Tests.” *Quarterly Journal of Economics*, 117(3), pp. 817-869.

[11] Chater, Nick, Steffen Huck and Roman Inderst (2010), "Consumer Decision-Making in Retail Investment Services: A Behavioral Economics Perspective", Report to the European Commission/SANCO.

[12] Cox, James C. (2004), “How to Identify Trust and Reciprocity,” *Games and Economic Behavior*, 46, 260-281

[13] Dufwenberg, Martin and Gneezy, Uri (2000). “Measuring Beliefs in an Experimental Lost Wallet Game,” *Games and Economic Behavior*, 30, pp. 163-182.

[14] Eagly, A.H. and M. Crowley (1986). “Gender and Helping Behavior: A Meta-Analytic Review of the Social Psychological Literature," *Psychological Bulletin*, 100, pp. 283-308.

[15] Eckel, Catherine C. and Philip J. Grossman (1998). “Are Women Less Selfish Than Men?: Evidence from Dictator Experiments," *Economic Journal*, 108, pp. 726-35.

[16] Ermisch, John and Diego Gambetta (2006). “People’s trust: the design of a survey-based experiment," *ISER Working Paper Series 2006-34*, Institute for Social and Economic Research.

[17] Fehr, Ernst (2009), “On the Economics and Biology of Trust”, *Journal of the European Economic Association*, 7, pp. 235-266.

[18] Fehr, Ernst and Urs Fischbacher (2004), “Third-Party Punishment and Social Norms.” *Evolution and Human Behavior*, 25, pp. 63-87.

[19] Fehr, Ernst and K.M. Schmidt (1999). “A theory of fairness, competition, and cooperation.” *Quarterly Journal of Economics*, 114 (3), pp. 817-868.

[20] Geanakoplos, John, David Pearce and Ennio Stacchetti (1989), “Psychological Games and Sequential Rationality,” *Games and Economic Behavior*, 1, pp. 60-79.

[21] Glaeser, Edward, David Laibson, Jose A. Scheinkman and Christine L. Soutter (2000), “Measuring Trust,” *Quarterly Journal of Economics* 115(3), 811-846.

[22] Gneezy, Uri (2005), “Deception: The role of consequences,” *American Economic Review*, March 2005, 384-394.

[23] Hung, Angela A., Clancy Noreen, Jeff Dominitiz, Eric Talley, Calude Berrebi and Farrukh Suvankulov (2008), "Investor and Industry Perspectives on Investment Advisers and Broker-Dealers", Technical Report, Rand Institute for Civil Justice.

[24] Inderst, Roman and Marco Ottaviani (2012), "Financial Advice," *Journal of Economic Literature*, 50(2): 494-512.

[25] Krupka, Erin L. and Roberto A. Weber (2013), “Identifying social norms using coordination games: Why does dictator game sharing vary?,” *Journal of the European Economic Association*, 11(3): 495-524.

[26] Rabin, Matthew (1993), “Incorporating Fairness into Game Theory and Economics.” *American Economic Review*, 83(5), pp. 1281-1302.

[27] Reiss, Michelle C., and Kaushik Mitra (1998). “The Effects of Individual Difference Factors on the Acceptability of Ethical and Unethical Workplace Behaviors, *Journal of Business Ethics*, 17(14), pp. 1581-93.

[28] Ross, Lee, Greene, D., and House, P. (1977), “The False Consensus Phenomenon: An Attributional Bias in Self-Perception and Social Perception Processes,” *Journal of Experimental Social Psychology*, 13(3), 279-301.

[29] Rousseau, Denise and Sim B. Sitkin and Ronald S. Burt and Colin Camerer (1998), “Introduction to Special Topic Forum: Not So Different After All: A Cross-Discipline View of Trust,” *The Academy of Management Review*, 23(3), pp. 393-404.

[30] Rustichini, Aldo and Marie-Claire Villeval (2012), “Moral Hypocrisy, Power and Social Preferences,” *GATE Working Paper No. 1216.*

[31] Schlag, Karl and Joel J. van der Weele (2013), “Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality,” *Theoretical Economics Letters*, 3(1), 38-42.

[32] Sapienza, Paola, Anna Toldra and Luigi Zingales (2007), “Understanding Trust,” NBER WP 13387

[33] Smith, Vernon L. (1976), “Experimental economics: Induced value theory.” *American Economic Review*, 66(2): 274-279.

Table 1: Experimental design

	Number of sessions	Explicit cheating notion question opt-out	Investment fee	Max belief pay	Obs
Initial study	4	No	0.50 euro	5 euro	122
Additional sessions	4	Yes	0.50 euro (2 sessions) 0.00 euro (2 sessions)	20 euro	306

Table 2: Descriptive statistics

	Mean	Std Dev	Min	Max	N
Male	0.46	0.499	0	1	420
Age	23.73	4.171	18	58	420
Math score	7.66	1.251	3	10	402
Inc<30K	0.29	0.455	0	1	391
30≤Inc<45	0.24	0.426	0	1	391
45≤Inc<70	0.25	0.431	0	1	391
70≤Inc<120	0.16	0.366	0	1	391
Inc≥120K	0.07	0.249	0	1	391
Risk aversion	5.71	2.193	1	10	417
Send decision (binary)	0.81	0.392	0	1	428
Send amount	4.31	3.232	0	10	428
Average return proportion	1.28	0.697	0	4.02	427
Average expected return proportion	1.27	0.637	0	4.02	425
Competitive values emphasis	0.62	0.196	0	1	410
Good values emphasis	0.76	0.149	0.17	1	404
Expected probability of not being cheated	0.42	0.232	0	1	427
Average proportion of non-cheaters	0.49	0.376	0	1	428
Own cheating notion					

Table 3: Proportion of participants in sessions who opt-out of reporting a cheating notion, restricted to sessions with explicit opt-out opportunities "this has nothing to do with being cheated," by send amount

	Send Amount										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	Obs
	Proportion who selected "this has nothing to do with cheating"										
Mean	0.20	0.18	0.17	0.15	0.13	0.13	0.13	0.13	0.14	0.13	306
Std. Error	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	
	Proportion who did not report a cheating notion for any reason										
Mean	0.23	0.21	0.21	0.17	0.15	0.15	0.15	0.16	0.17	0.17	306
Std. Error	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	

Notes: [1] In sessions with an explicit "opt-out" possibility participants could refrain from specifying an explicit personal cheating notion and instead respond either "I don't know" or "this has nothing to do with cheating." [2] The top row of Table 3 presents the proportion of participants who chose "this has nothing to do with cheating," while the lower row presents the proportion of participants who chose either of these two "opt-outs" or left the question entirely blank.

Table 4: Proportion of participants whose cheating notions are consistent with various definitions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
WPROI	0.16	0.15	0.18	0.14	0.16	0.12	0.12	0.10	0.11	0.14
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
SPROI	0.31	0.29	0.36	0.31	0.29	0.24	0.20	0.18	0.23	0.28
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Inequality Aversion	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)	0.10 (0.02)	0.30 (0.02)	0.34 (0.02)
ES	0.33	0.32	0.30	0.24	0.29	0.32	0.33	0.29	0.30	0.34
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Obs	359	365	365	375	381	382	383	378	377	377

Notes: [1] The table reports the proportion of respondents reporting two prominent classes of cheating notions. Classifications are not mutually exclusive so that, e.g., the same cheating notion can be labeled as consistent with both SPROI and Inequality aversion. [2] Standard errors appear in parentheses. [3] A weakly positive return on investment (WPROI) cheating notion entails reporting exactly the send amount (s) as one's cheating threshold in sessions without a sending fee. In sessions with a sending fee (€0.50), WPROI includes both participants who report s and participants who report s + €0.50 as their cheating threshold. [4] "SPROI" is a more generous definition of PROI taking into account a reasonable interest rate, r = 10%. We compute the send amount plus the sending fee, where applicable, and multiply this by 1+r to get an "exact SPROI" definition. To be as generous as possible to this notion, and to account for the fact that experimental participants typically have a well-known predilection to state whole-number values, we then calculate the least integer greater than this exact value, denoted by ceiling("exact SPROI"). For each send amount, s, We label as SPROI all cheating thresholds falling within the interval with integer end-points: [s, ceiling("exact SPROI")]. [5] "Inequality Aversion" refers to a cheating notion which requires equal monetary outcomes, and we label a cheating notion as consistent with inequality aversion if it lies within the smallest closed interval with integer endpoints containing this outcome. As an example, suppose there is a €0.50 sending fee and consider s = 1. The total surplus in this case is 10.50 – 0.50 +– 1 + 8.05 = 17.05, and half of this surplus is 8.525. Any cheating notion in the interval [8, 9] would therefore be labeled as consistent with inequality aversion. [6] An "Equal-split" (ES) cheating notion entails a cheating threshold of half of the entire amount allocated to the receiver. As with SPROI above, to account for participants' predilection for whole numbers, the definition of ES for each send amount, s, includes all cheating thresholds falling within the smallest interval with whole-number end-points containing a precisely-equal split of the receivers' total earnings: i.e., $\frac{f(s)}{2} \in$ [n, n+1]. For example, if a sender sends s = 3, a receiver receives f(s) = 11.30, and $\frac{f(s)}{2}$ = 5.65. Consequently, ES for s = 3 would include all cheating thresholds within the interval [5, 6].

Table 5: Determinants of cheating notions

Dependent variable = Own cheating notion										
Cooperative values	Competitive values	€ sent	(€ sent)^2	Male	Age	Math score	Risk aversion	Cons	Obs	Individuals
-2.55**	1.63**	1.07***	-0.02***	-0.47	0.00	-0.02	-0.11	3.55***	3496	354
(1.09)	(0.64)	(0.07)	(0.01)	(0.43)	(0.03)	(0.11)	(0.08)	(1.31)		

Notes: [1] Estimates are from an individual-level random effects regression model. [2] Variables present in the regression, but omitted for readability: full set of income dummies; dummy for sessions with no investment fee; dummy for sessions comprising the initial study. None of these variables had significant coefficients. [3] Robust standard errors, clustered by session, appear in parentheses.

Table 6: Intentional cheating by send amount

					Send Amount					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Own cheating notion	-0.08**	-0.13***	-0.08*	-0.07***	-0.04	-0.04*	-0.04	-0.05**	-0.03*	-0.04*
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
Est. others' cheating notion	0.22***	0.21***	0.23***	0.22***	0.17***	0.15***	0.16***	0.17***	0.12***	0.14***
	(0.04)	(0.06)	(0.05)	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
Male	0.07	-0.02	0.18	0.06	0.03	-0.05	-0.16	-0.07	0.01	-0.06
	(0.07)	(0.16)	(0.12)	(0.14)	(0.20)	(0.15)	(0.14)	(0.13)	(0.13)	(0.18)
Age	-0.03	-0.04***	-0.03***	-0.02**	-0.02**	-0.03**	-0.01	0.01	-0.01	-0.01
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Math score	-0.04	-0.03	0.04	0.05	-0.04	-0.03	-0.06***	-0.06	-0.08**	-0.03
	(0.06)	(0.05)	(0.04)	(0.05)	(0.05)	(0.09)	(0.02)	(0.06)	(0.04)	(0.04)
Risk aversion	-0.00	0.03	0.01	-0.00	-0.02	-0.02	0.00	-0.01	-0.02	-0.07***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.02)
30≤ Inc	-0.02	0.18	0.24**	0.22	0.09	0.25*	0.50*	0.06	0.10	0.16
	(0.19)	(0.16)	(0.12)	(0.21)	(0.23)	(0.15)	(0.26)	(0.19)	(0.12)	(0.21)
45≤	0.12	0.01	0.06	0.10	0.29*	0.23***	0.43	0.24**	0.12	0.15
	(0.16)	(0.08)	(0.13)	(0.17)	(0.17)	(0.07)	(0.28)	(0.12)	(0.14)	(0.20)
70≤ Inc	0.17	0.17	0.07	-0.05	0.33*	0.41*	0.58**	0.70***	0.04	0.16
	(0.33)	(0.18)	(0.21)	(0.18)	(0.19)	(0.22)	(0.24)	(0.16)	(0.20)	(0.33)
Inc ≥120	0.00	-0.21	-0.07	0.01	-0.51	0.02	-0.21	-0.44	-0.04	-0.56*
	(0.35)	(0.28)	(0.21)	(0.20)	(0.32)	(0.28)	(0.40)	(0.31)	(0.29)	(0.33)
Constant	0.45	0.85	-0.69	-1.02*	-0.07	-0.10	-0.76*	-0.97	-0.05	-0.42
	(0.76)	(0.52)	(0.52)	(0.60)	(0.41)	(0.79)	(0.41)	(0.81)	(0.70)	(0.63)
Obs	369	366	366	369	371	370	371	369	366	366

Notes: [1] Each column presents estimates from a Probit model, with the (binary) dependent variable being "receiver intentionally cheats if sent relevant amount." Intentional cheating is defined by sending back strictly less than the receiver estimated senders needed back in order to not feel cheated. This threshold amount is also inserted as a control in each estimate by the variable "Est. others' cheating notion." [3] Robust standard errors, clustered by session, in parentheses. *** = significant at 1%, ** = significant ay 5%, * = significant at 10%. [4] Math score is individual's self-reported score on required math exams taken during the final year of high school in Italy. [5] Income variables refer to self-reported annual family income from all sources, in thousands of euros, net of taxes. The excluded category is "below 30 thousand euros annually". [6] Observations vary over columns because not all participants reported a cheating notion for every send amount. This is discussed in the text. Additionally, we do not have demographics for all participants.

Table 7: Senders' decisions, Heckman estimates

	Main equation (1)	Selection equation (2)
Expected probability of not being cheated	2.76**	0.57
	(1.38)	(0.65)
Expected return from trusting	1.34***	0.28**
	(0.45)	(0.12)
(Probability of not being cheated) x(Expected return from trusting)	-1.57*	-0.07
	(0.85)	(0.46)
Low fee (dummy)	--	0.68***
		(0.09)
Age	0.11***	0.00
	(0.03)	(0.02)
Male	0.36	0.35**
	(0.32)	(0.14)
Math score	-0.00	0.12***
	(0.09)	(0.04)
Risk aversion	-0.14***	0.04
	(0.05)	(0.03)
Altruism	0.03	0.04
	(0.12)	(0.04)
30≤ Income <45	-0.29	0.13
	(0.42)	(0.25)
45≤ Income <70	-0.22	-0.04
	(0.59)	(0.23)
70≤ Income <120	-0.62**	-0.08
	(0.29)	(0.13)
Income ≥120	-0.63	0.74*
	(0.70)	(0.40)
Constant	1.45	-1.62***
	(2.16)	(0.61)
Obs	350	350
Mills Ratio	-.86	
	(0.36)	

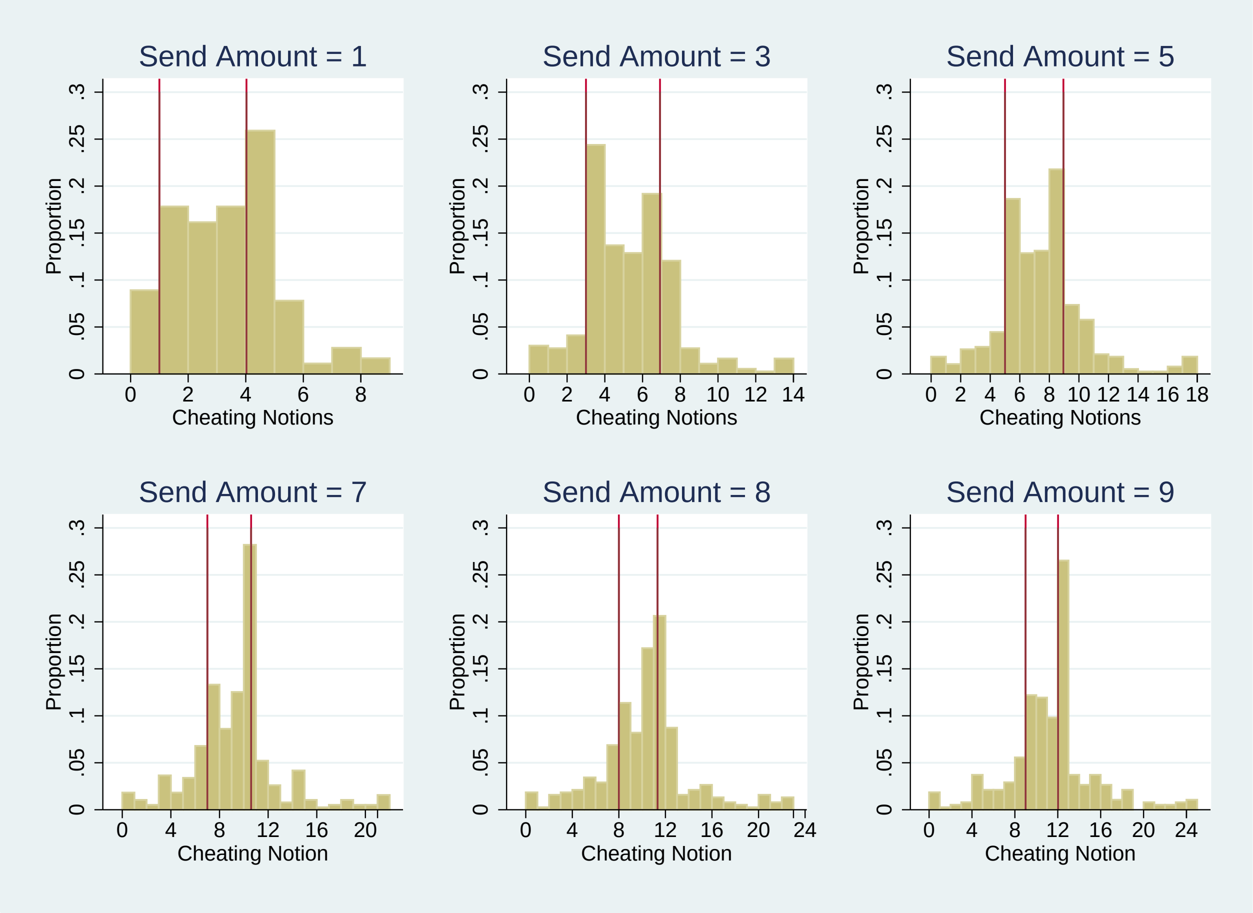
Notes: [1] Robust standard errors, clustered by session, appear in parentheses. [2] *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [3] For the Heckman model (cols 1-2): the dependent variable in the selection equation takes the value of 1 if the sender sends a positive amount and 0 otherwise; the dependent variable in the main equation is *how much* the sender sends. [4] The exclusion restriction for the selection equation consists of a dummy for "Low fee" sessions, a dummy taking the value of one if the observation came from a session where senders were charged nothing to send a positive amount, and 0 if the observation came from a session where senders were charged € 0.50 to send a positive amount [5] "Expected probability of not being cheated" is our measure of participants' subjective beliefs about not being cheated, described in the text. [6] "Expected return from trusting" is the participant's estimate of the proportion of money *sent* that receivers will return, averaged over all 10 possible send amounts. [7] "Risk aversion" is an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism in a separate, unrelated, experiment. This variable takes values from 1 (risk loving) to 10 (very risk averse). [8] Altruism is how much emphasis participants' parents placed on the value "help others" during their upbringing. [9] Income variables refer to (self-reported) annual family income from all sources, in thousands of euros, net of taxes. The lowest category is excluded: "below 30 thousand euros".

Table 8: Sensitivity of amounts returned to cheating notions by decision to cheat, Heckman models

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Conditional on not cheating (euros returned >= estimated others' cheating notion)										
Est. others' cheating notion	1.17***	1.02***	0.97***	1.19***	1.07***	0.95***	0.88***	0.89***	1.09***	1.02***
	(0.17)	(0.13)	(0.14)	(0.25)	(0.15)	(0.11)	(0.16)	(0.13)	(0.22)	(0.13)
Constant	3.83**	4.98**	3.54**	4.68*	5.06**	4.88***	5.96***	4.97**	8.15**	9.26***
	(1.95)	(2.25)	(1.73)	(2.78)	(2.39)	(1.85)	(2.10)	(2.00)	(4.06)	(3.02)
Demographics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs	311	319	320	328	333	334	335	332	329	329
Wald test: *est. others' cheating notion* coefficient = 1 (p-value)										
	0.32	0.86	0.84	0.43	0.63	0.66	0.43	0.39	0.67	0.84
Conditional on cheating (euros returned < estimated others' cheating notion)										
Est. others' cheating notion	0.42***	0.37***	0.57***	0.38***	0.44***	0.43***	0.49***	0.58***	0.63***	0.53***
	(0.07)	(0.06)	(0.10)	(0.10)	(0.10)	(0.09)	(0.13)	(0.12)	(0.14)	(0.12)
Constant	-0.16	0.93	0.18	0.95	1.68	0.03	1.09	0.09	-0.38	-0.01
	(0.92)	(1.02)	(1.51)	(1.92)	(1.91)	(2.01)	(2.87)	(3.02)	(3.36)	(3.31)
Demographics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs	311	319	320	328	333	334	335	332	329	329
Wald test: *est. others' cheating notion* coefficient = 0.5 (p-value)										
	0.23	0.04	0.47	0.24	0.52	0.43	0.95	0.52	0.34	0.79

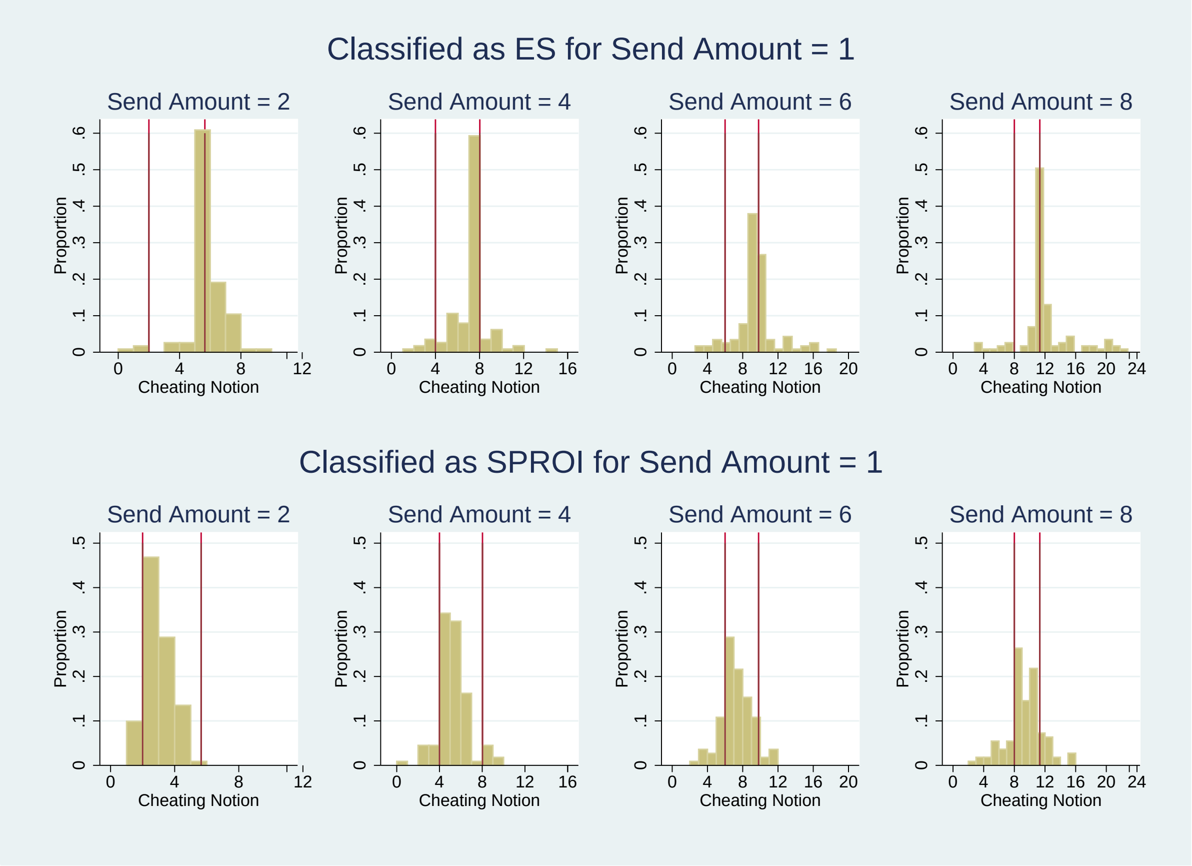
Notes: [1] Standard errors in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents a Heckman model estimate using as its exclusion restriction participants' own cheating notions. [3] The dependent variable in column i is the amount a participant will send back if the sender sends i euros, i=1,…,10. [4] The reported independent variables in column *i* are: "Est others' cheating notion" is each participant's estimate of the minimum amount of money a sender would need back in order to not feel cheated when the sender sends *i* euros, i=1,…,10. [5] Each estimate includes our standard set of demographic controls, omitted for readability from the table. These controls are: gender, age, math score, family income and risk aversion.

Figure 1: Own Cheating Notions, Histograms



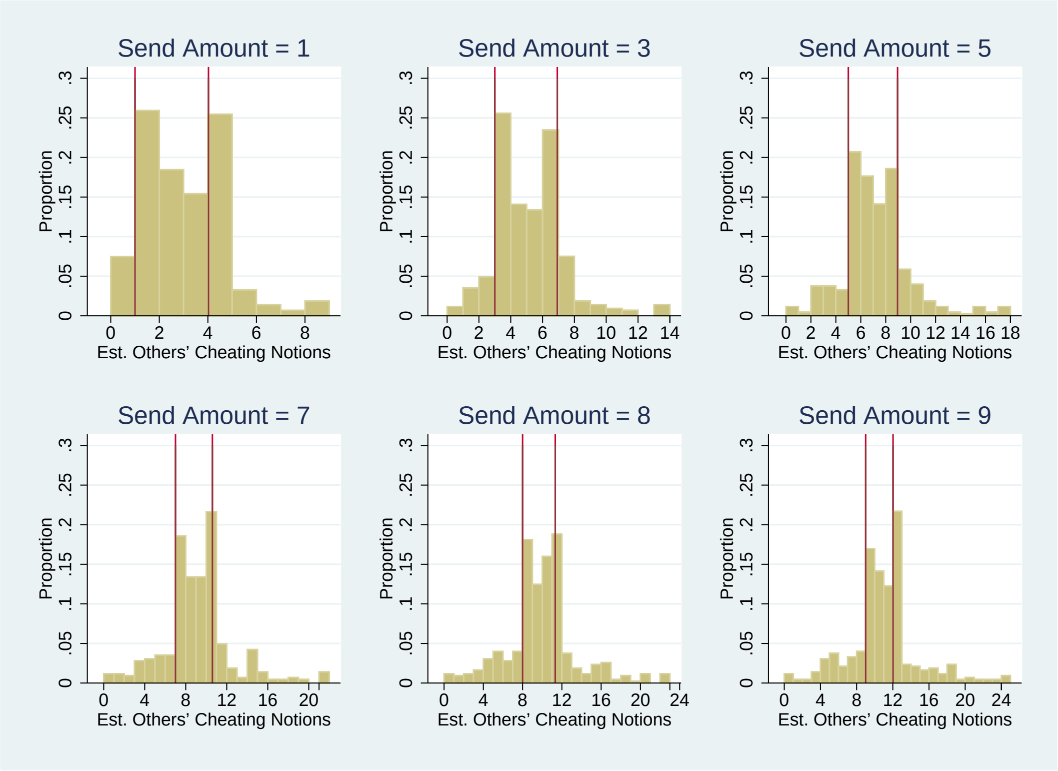
Notes: [1] The figure reports histograms of participants' personal cheating notions for several send amounts. Histograms for the omitted send amounts look similar. [2] Each histogram is overlaid with two vertical bars. The first bar is the send amount, and corresponds to a WPROI cheating definition; the second bar occurs at half of the total amount receivers' receive and corresponds to an ES cheating definition.

Figure 2: Consistency of Cheating Notions across Send Amounts.



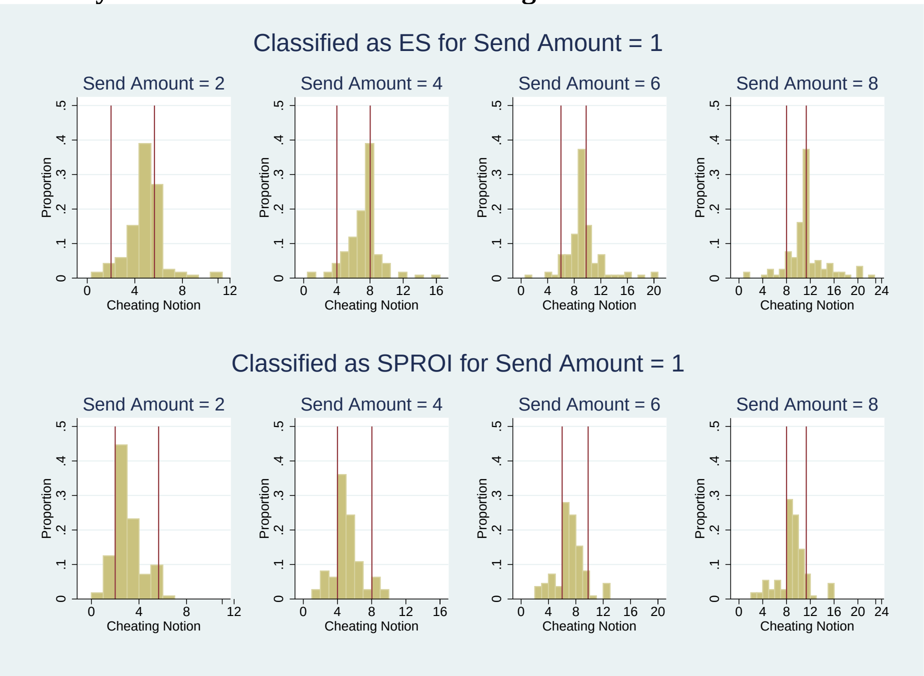
Notes: [1] The figure restricts attention to participants whose cheating notions were consistent with ES (top row) or SPROI (bottom row) conditional on a send amount of 1, and presents histograms of these participants' cheating notions for various other send amounts. Omitted histograms look similar. [2] Each histogram is overlaid with two vertical bars. The first bar is the send amount, and corresponds to a WPROI cheating definition; the second bar occurs at half of the total amount receivers' receive and corresponds to an ES cheating definition.

Figure 3: Participants' Beliefs about Others' Cheating Notions, Histograms



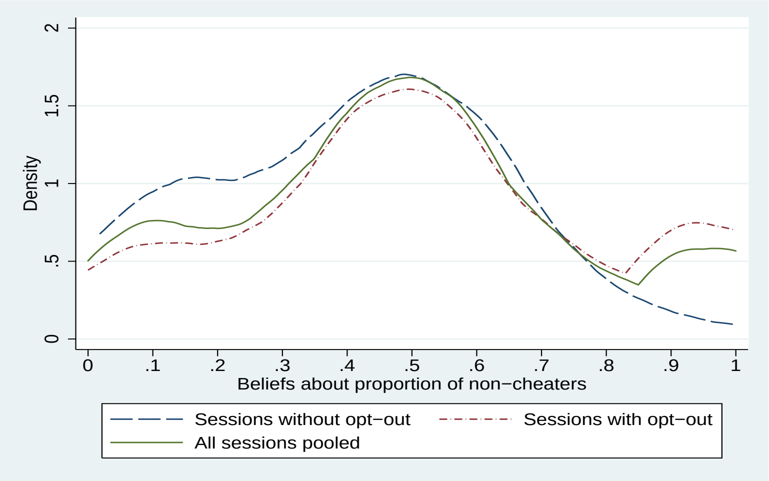
Notes: [1] The figure reports histograms of participants' beliefs about other participants' personal cheating notions for several send amounts. Histograms for the omitted send amounts look similar. [2] Each histogram is overlaid with two vertical bars. The first bar is the send amount, and corresponds to a WPROI cheating definition; the second bar occurs at half of the total amount receivers' receive and corresponds to an ES cheating definition.

Figure 4: Consistency of Beliefs about Others' Cheating Notions across Send Amounts.



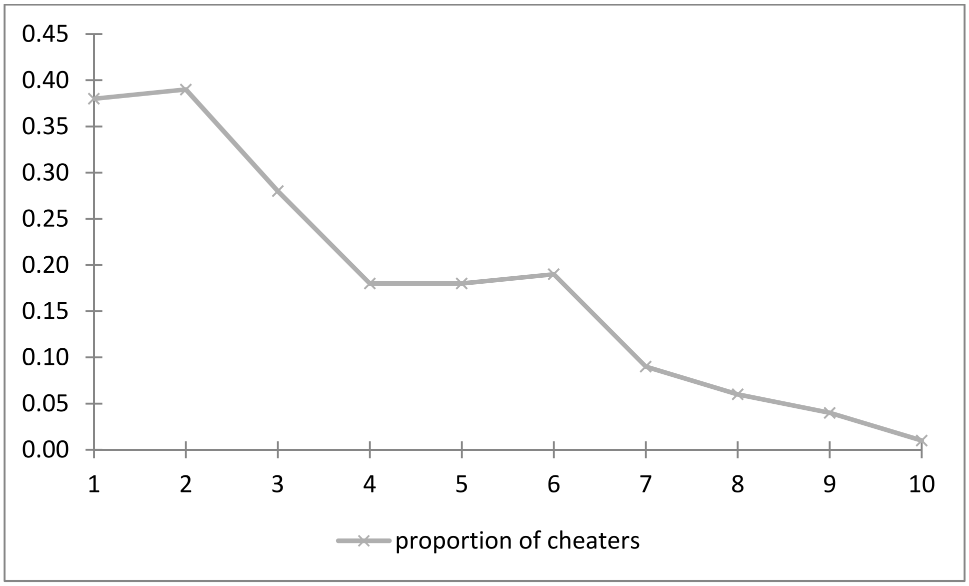
Notes: [1] The figure restricts attention to participants whose cheating notions were consistent with ES (top row) or SPROI (bottom row) conditional on a send amount of 1, and presents histograms of these participants' beliefs about others' cheating notions for various other send amounts. Omitted histograms look similar. [2] Each histogram is overlaid with two vertical bars. The first bar is the send amount, and corresponds to a WPROI cheating definition; the second bar occurs at half of the total amount receivers' receive and corresponds to an ES cheating definition.

Figure 5, Beliefs about the probability of not being cheated



Notes: Observations in the sessions with opt-out (short-dash line) are restricted to individuals who have a cheating notion for every possible amount a sender could send. This is to ensure our summary measure of beliefs about the probability of being cheated is well-defined. Thus the density plot for the additional sessions is based on 207 (out of 306) observations.

Figure 6: Proportion of cheaters by send amount



Notes: The figure reports the proportion of cheaters (y-axis), after partialling out the effect of expectations of others' cheating notions, for each possible send amount (x-axis).

Not for publication

Appendix I: Robustness Checks

A Additional Robustness check treatments

In addition to our main experiment described in Appendix II, two further treatments were conducted for robustness. First of all, to check whether there is something peculiar about the on-line environment driving our results or whether paying only 10 percent of participants provides incentives that are too weak, we ran two sessions in the laboratory where 100 percent of participants were paid. As a second robustness exercise, we conducted sessions in which our direct cheating notion question was omitted and replaced with a series of questions asking participants how they would feel about various possible outcomes in the trust game from the point of view of the sender. The purpose of this latter treatment is to address the concern that our direct cheating notion question might prime participants to associate cheating with the trust game.

A.1 In-lab sessions

In total, 36 individuals took part in two sessions conducted in the experimental laboratory at the Einaudi Institute for Economics and Finance in Rome, Italy. Participants were recruited from the same subject pool as were the on-line sessions. There was no overlap in actual participants—i.e., no participant took part in both an on-line session and an in-lab session. All in-lab participants were paid based on their choices in the experiment and the accuracy of the their reported beliefs.

Apart from taking place in the laboratory, the design of this treatment and the materials used were exactly the same as the on-line treatments. Participants simply completed the on-line experiment in the laboratory. All sessions of the in-lab experiment allowed participants to opt out of specifying a cheating notion by selecting one of two responses: "I don't know" or "this has nothing to do with cheating." Neither session featured a fee to send a positive amount.

In Table A1 we report summary statistics for both the in-lab and most comparable on-line sessions. Receivers' behavior does not change much across these two environments: average return proportions and the propensity to intentionally cheat are all quite similar. Beliefs about these return proportions and the likelihood of being cheated are also quite similar across the two environments. On the other hand, in-lab senders were slightly more likely to send a positive amount than their on-line counterparts, raising the average amount sent by in-lab senders. However, conditional on sending a positive amount average send amounts were again quite similar: 5.36 in on-line low fee sessions; 5.43 in the laboratory; with standard errors 0.25 and 0.44, respectively.

In terms of cheating notions, the picture is also quite similar in the lab and on-line experiments: the vast majority of participants have a cheating notion for all possible send amounts (Table A2); the vast majority have a cheating notion *at least as demanding as* WPROI (TableA3). Considering the proportion of participants whose cheating notions are consistent with various definitions (Table A4), we again see that WPROI describes a small minority of participants, while a similar but relaxed notion, SPROI, describes a substantial minority of participants for most send amounts, as does an ES rule: over all send amounts, these two rules each account for about 27%-29% of participants' reported cheating notion. We also, again, find that literal inequality aversion fits very few participants' definitions of cheating. We find the same patterns when considering beliefs about others' cheating notions (Table A5), which is also consistent with our on-line findings.

Considering next the relationship between second-order beliefs and cheating notions and related beliefs, the in-lab environment delivers similar patterns as those found in the on-line environment. Own cheating notions are again highly predictive of return proportion beliefs (Table A6). In-lab beliefs about others' cheating notions are highly predictive of in-lab second-order beliefs (Table A7). As in the on-line data, own cheating notions are negatively related to intentional cheating while expectations about others' cheating notions are positively related to intentional cheating (Table A8).

In Table A9, we replicate the pattern suggesting that beliefs about others' cheating notions function as thresholds for those who refrain from cheating. Because we have many fewer observations here, to show this we take a more straightforward approach and do not model selection explicitly. Instead, we simply split the data into those who refrain from intentional cheating (top panel) and those who intentionally cheat (bottom panel) and run simple univariate OLS regressions of return amounts on beliefs about others' cheating notions. We find that, just as in the main data, for those who refrain from intentionally cheating return amounts vary essentially one-to-one with cheating notion beliefs for most send amounts. For those who intentionally cheat, return amounts are consistently much less sensitive to cheating notion beliefs which is, again, consistent what we find in the on-line data.

Considering the sender's side of the exchange, next we consider how send amounts vary with cheating and monetary return beliefs (Table A10). Because we have few observations and lack the exogenous variation in senders' incentives which we exploited in the analysis of our on-line data, we account for selection into sending a positive amount here by estimating a Tobit model rather than a Heckman model. The results paint a picture qualitatively similar to the on-line data: amounts sent vary positively and significantly with both expected (lack of) cheating and expected return.

A.2 Treatments without cheating notion question

We also conducted (on-line) sessions of a treatment in which we dropped our direct cheating notion question and replaced it with a section where participants were asked to indicate how they would feel, as a sender, about various send/return amount scenarios. In total, 170 participants took part in this treatment. As with the main study, ten percent of participants were randomly chosen to be paid their experimental earnings.

To keep the number of individual questions reasonable, we selected three common send amounts—$S = 1, 5$ and 10—and, for each of these, asked participants how they would "feel" if the receiver returned four specific amounts: $0, \frac{S}{2}$, S and $\frac{f(S)}{2}$. These send/return scenarios were chosen to line up with the cheating notions common in the data from our main study. In terms of feelings, for each send/return amount scenario participants were asked to select exactly two options from a list of several options that best described how they would feel if the scenario were realized. The list of options included positive evaluations ("[the receiver] was generous," "[the receiver] treated me fairly"), neutral evaluations ("[the receiver] was intelligent," "I have no particular opinion of [the receiver's] behavior") and negative evaluations ("[the receiver] cheated me," "[the receiver] disappointed me"). A free-form response option was also available.

To compare the qualitative data we have in this treatment with data from our main sessions, for each send/return scenario investigated in this treatment we calculate the proportion of participants in our main treatment who would feel cheated according to their own reported cheating notions. We compare this proportion to the proportion of respondents in the "feelings" treatment reporting feeling "disappointed" or "cheated." To maximize comparability, from our main treatment data we use only sessions where participants were allowed to opt out of specifying a cheating notion. We find a strong positive relationship between the proportion of participants expressing negative feelings in particular scenarios and the implied proportion of participants feeling cheated in those scenarios in the data from the main treatment (Figure A1). We interpret this as support for the view that trust game participants have well-defined cheating notions and evidence against the view that the cheating notions they report can be mainly attributed to priming.

A.2.1 Evidence on receivers' motivations

In sessions without a direct cheating notion question, at the end of the experiment we added a section in which participants were asked to descibe the rationale they used, if any, for deciding how much to return in the role of receiver. Participants were asked:

> *Describe, in general, how you arrived at your decisions concerning how much to return when you played role B [receiver] for each amount A could have sent you*

Participants could select among four pre-programmed options, or, if none on the list suited them they could select "other" and specify their own rationale. Three of the four pre-programmed responses were meant to capture positive reciprocity, ("the more A [the sender] sent, the more I returned in order to reward nice behavior"); negative reciprocity ("the less A [the sender] sent, the less I returned, in order to punish bad behavior"); vulnerability ("the more A [the sender] sent, the more I returned in order to compensate A [the sender] for being at the mercy of my actions"). The fourth pre-programmed option was essentially a decline to state option ("I did not have any particular rationale in mind.").

Table A11 presents the results. Overall, 83 percent of participants selected one of the four pre-programmed option. The modal response, selected by 42 percent of participants, was that receivers return more when senders send more to compensate senders for their vulnerability. The second most common response reflected positive reciprocity. Almost nobody (6 percent) selected negative reciprocity as their primary rationale, while a similarly low percentage selected the pre-programmed decline to state option (6 percent).

B Robustness checks on beliefs

A common concern whenever beliefs are elicited is the extent to which the elicitation mechanism itself colors reported beliefs. Monetary incentives meant to ensure that participants report beliefs truthfully may give rise to other potential confounds, such as hedging motives: by shading reported beliefs toward bad outcomes, individuals may reduce the variance of their experimental earnings. On the other hand, monetary incentives that are too weak can allow reported beliefs to be non-truthful for various reasons. In particular, one may worry that the significant correlation between estimates of others' cheating notions and own return amounts arises because of a tendency for participants to ex-post rationalize their receiver strategies: by reporting believing that whatever they return is enough to not cheat others, participants can maintain a positive moral self-image.

First we consider ex-post rationalization. If ex-post rationalization is driving beliefs about others' cheating notions, then quadrupling the incentives for belief accuracy in the additional sessions should make this motive less relevant. Evidence of ex-post rationalization would be a consistently smaller correlation between return amounts and beliefs about others' cheating notions in the "high belief pay" sessions.

As a simple test for ex-post rationalization, Table A12 (panel A) presents panel regressions of beliefs about others' cheating notions as a function of return amounts incorporating a dummy for high belief pay and an interaction with return amounts. The coefficient of interest is on the interaction between high belief pay and return amount: if ex-post rationalization is important when belief pay is low, and diminished for high belief pay, we would expect this coefficient to be consistently negative and significant. Instead, the estimated coefficient on the interaction term is positive and marginally significant providing evidence against ex-post rationalization. Adding our standard set of demographics does not change the results. Moreover, restricting to the subset of observations where the receiver does not intentionally cheat—where the ex-post rationalization argument has the most bite—changes nothing qualitatively. We omit these last two robustness checks to save space, but they are available on request. It should also be noted that variation in belief pay could not have directly affected receivers' actions, since participants did not know there would be a belief elicitation section until after they had submitted their strategies.

Next, consider hedging motives. As a concrete example, consider a sender who has chosen to send 10 euros. If the sender believes the receiver is trustworthy and reports this belief, then in the good state of the world where the receiver *is* trustworthy, the sender earns a lot—both beliefs and actions pay off. However, in the bad state of the world, say, where the receiver returns nothing, the sender loses quite a lot—neither actions nor beliefs pay off. By shading reported beliefs downward—towards a higher likelihood of an untrustworthy sender—the sender can shift some earnings out of the good state of the world into the bad state of the world, reducing earnings variance, i.e., risk.

To test for hedging motives in beliefs, we estimate participants' stated beliefs about the amount of money receivers will return for each possible send amount. We present panel regressions, where we control for whether a sender actually chose to send a particular amount, risk aversion and an interaction between these two variables. Since hedging motives can only (literally) apply to the send amount a sender actually chooses, one measure of the hedging motive is the coefficient on the dummy for actually-chosen send amounts. A secondary prediction is that more risk averse individuals care about hedging more, so the interaction term should be negative. Table A12 (panel B) presents our estimates, which provide no support for the importance of hedging. In fact, contrary to hedging motives, reported beliefs about return amounts are marginally significantly *higher* for the amount a sender actually chose to send as evidenced by the coefficient on "Chosen send amount." Risk aversion plays no significant role. Controlling for demographics and/or the level of belief pay does not change anything qualitatively, so we omit these specifications.

C Additional Robustness checks on cheating notions

One additional concern with cheating notions is that they may be (reverse) caused by beliefs. Although priming is not an issue here, as we elicited beliefs after cheating notions, one explanation for the strong correlation between cheating notions and senders' (first-order) beliefs about receivers' return amounts could be that that individuals simply report how much they expect back from receivers as their cheating notion. One reason this could happen is through an individual's desire to maintain a positive self-image and to avoid appearing, to themselves or to the experimenters, as "foolish" for allowing themselves to be cheated. To be clear, if senders expect not to be cheated and hence their cheating notion affects their reported beliefs, that is fine for our purposes. However, if participants first form beliefs about how much receivers will return and then report this belief as their cheating notion because of, e.g., a desire to not appear like a "sucker," then this calls into question the informativeness of the reported cheating notion.

In the latter case, it seems likely that such processes would affect reported cheating notions much more strongly for situations which could *actually* occur—i.e., for the one send amount an individual actually chooses. For concreteness, suppose an individual chooses to send $s = 3$ in the role of sender. Since this is an event that may actually occur, when asked about his or her cheating notion for $s = 3$ an individual may report his or her belief about how much the receiver will return instead of his or her cheating notion in order to avoid the looking like a sucker if the event actually occurs, particularly problem if the return belief is greater than the cheating notion. Such a process would tend to inflate reported cheating notions and possibly overstate the correlation between cheating notions and first-order beliefs. However, for all other send amounts ($s = 1, 2, 4, \dots, 10$), since they cannot actually occur, such processes should have little effect on cheating notions or their relationship to first-order beliefs.

To test for this effect, we report in Table A13 the results of ten separate regressions—one for each send amount—using an individual's personal cheating notion as the dependent variable. On the right hand side, we include an individual's beliefs about the amount the receiver will return, a dummy indicating whether the individual chose to send the amount listed in the column heading and an interaction between these two variables. We control for our usual set of demographics, but as they have little explanatory power here we do not report them for ease of exposition.

We find that whether an individual actually chooses a particular send amount has no consistent effect on his or her reported cheating notion: half of the estimated coefficients on *Chosen send amount* are positive, half are negative, and only one out of the ten coefficents is significant at conventional levels. Similarly, whether an amount was actually chosen has no consistent effect on the relationship between beliefs and cheating notions: five of the ten coefficients on the interaction between beliefs and cheating notions are positive, the other five are negative and only one out of the ten is statistically significant. Considered together, our results provide little evidence for cheating notions being reverse-caused by beliefs because, e.g., participants want to avoid looking like a sucker.

D Cheating notions and guilt aversion theory

In this section we test for the conjectured correlations between i) own cheating notions and beliefs about receivers' actions; and ii) beliefs about others' cheating notions and receivers' second order beliefs.

In Table A14 we report ten separate regressions—one for each send amount—using individuals' first-order beliefs about the amount receivers will return as the dependent variable and, as the main explanatory variable, an individual's own personal cheating notion. We control for available demographics and relevant experimental design features. In this latter category, we include a dummy for whether there was a sending fee in the session as this might factor into a sender's definition of return on investment. As a simple check on whether the reported beliefs are true beliefs, or rather whether the relationship between beliefs and cheating notions is driven by nuisance factors (e.g., ex-post rationalization), we include a dummy indicating sessions where we *quadrupled* belief elicitation incentives as well as an interaction term between this dummy and own cheating notions. The main lesson from this exercise is that one's own cheating notion is consistently a highly significant predictor of senders' first-order beliefs. The strength of the relationship is large in magnitude as well: a one euro increase in personal cheating notions translates into a roughly 50 cent increase in the amount senders believe receivers will return. Examining the coefficient on the interaction between cheating notions and belief elicitation incentives, we find that much stronger incentives have no consistent impact on this relationship and that, moreover, the impact is almost never significant. These patterns suggest that reported beliefs are true beliefs. Finally, it is worth noting that demographics have little explanatory power with one exception: gender. Male participants consistently expect about 40 to 50 cents less back from receivers than female participants.

In Table A15, we estimate receiver's second-order beliefs as a function of their (first-order) beliefs about others' cheating notions. As before, we control for available demographics, relevant experimental design features, beliefs incentives and an interaction between beliefs incentives and reported beliefs about others' cheating notions. We find that beliefs about others' cheating notions are always highly significant predictors of second-order beliefs and that this relationship is also large in magnitude: a one-euro increase in beliefs about others' cheating notions translates into a 34 to 83 cent increase in second-order beliefs with an average increase, over all ten send amounts, of about 60 cents. Strengthened belief incentives, again, have no consistent impact on this relationship and, moreover, their effect is almost never significant at conventional levels. Demographics play a slightly larger role here: being male or having more mathematical ability tends to lower second-order beliefs; being older tends to raise them. The main lesson from Table 6, however, is that beliefs about others' cheating notions exhibit a strong positive relationship with second-order beliefs.

In Table A16 we test whether second-order beliefs contain predictive power for receivers' behavior beyond what is contained in cheating notion beliefs. To answer this question, we regress the amount receivers return on both their cheating notion beliefs and their second-order beliefs as well as available demographics.[1,2] We do this for each possible send amount, $s = 1, \ldots, 10$, separately. We find that, controlling for cheating notion beliefs, the relationship between receivers' behavior and their second-order beliefs is typically non-significant, e.g., achieving a 5% significance level for only one send amount. On the other hand, for eight out of our ten regressions beliefs about others' cheating notions are highly significant predictors of receivers' behavior. The relationship is large in magnitude as well: taking the average over all ten estimated coefficients, a one-euro increase in receivers' beliefs about senders' cheating notions is associated with a 26.6 euro-cent increase in the amount a receiver returns.[3]

[1] We verify that collinearity does not bias the results by computing the variance inflation factor for second-order beliefs and cheating notion beliefs. The computed value for either regressor is never above 2, whereas as a rule of thumb it would take a value above 10 to raise concerns about collinearity.

[2] Because our demographic controls are almost never significant predictors of behavior—the exceptions being the two highest income category dummies—we omit them for readability.

[3] We lose observations by controlling for demographics. However, nothing changes qualitatively when demographic controls are omitted. The coefficients on second-order beliefs are still generally non-significant and small in magnitude.

Table A1: Comparison of behavior in the lab and on-line, summary statistics

	Send > 0	Send amount	Return proportion	Return proportion (belief)	Proportion of non-cheaters	Proportion of non-cheaters (belief)
			In-lab sessions			
	0.97	5.28	1.25	1.36	0.43	0.56
	(0.03)	(0.45)	(0.10)	(0.10)	(0.06)	(0.03)
Obs	36	36	36	36	36	36
			On-line low fee sessions			
	0.90	4.83	1.28	1.22	0.53	0.53
	(0.03)	(0.26)	(0.06)	(0.06)	(0.03)	(0.02)
Obs	150	150	149	148	150	135

Table A2: Proportion of participants with a cheating notion, in-lab sessions

	Send amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Proportion w/ cheating notion	0.72	0.86	0.83	0.92	0.94	0.94	0.97	0.97	0.97	0.97
	(0.08)	(0.06)	(0.06)	(0.05)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Obs	36	36	36	36	36	36	36	36	36	36

Notes: [1] Raw proportions reported. [2] Standard errors appear in parentheses

Table A3: Proportion of participants who would feel cheated by (return amount)<(send amount), in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Proportion w/ (cheating notion) ≥ (send amt)	0.88	0.84	0.97	0.94	0.94	0.97	0.89	0.83	0.83	0.86
	(0.06)	(0.07)	(0.03)	(0.04)	(0.04)	(0.03)	(0.05)	(0.06)	(0.06)	(0.06)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Reported proportions are conditional on specifying a cheating notion. [2] Standard errors appear in parentheses

Table A4: Proportion of participants cheating notions consistent with various definitions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
WPROI	0.08	0.10	0.30	0.18	0.18	0.12	0.20	0.14	0.11	0.23
	(0.05)	(0.05)	(0.09)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)	(0.05)	(0.07)
SPROI	0.15	0.16	0.50	0.39	0.29	0.32	0.29	0.29	0.23	0.31
	(0.07)	(0.07)	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.08)
Inequality Aversion	0	0.03	0	0.03	0.06	0.03	0	0.03	0.23	0.37
	--	(0.03)	--	(0.03)	(0.04)	(0.03)	--	(0.03)	(0.07)	(0.08)
ES	0.35	0.23	0.20	0.18	0.32	0.32	0.34	0.23	0.23	0.37
	(0.10)	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Reported proportions are conditional on specifying a cheating notion. Classifications are not mutually exclusive so that, e.g., the same cheating notion can be labeled as consistent with both SPROI and Inequality aversion. [2] Standard errors are in parentheses. [3] A weakly positive return on investment (WPROI) cheating notion entails reporting exactly the send amount (s) as one's cheating threshold in sessions without a sending fee. [4] "SPROI" is a more generous definition of WPROI taking into account a reasonable interest rate, r = 10%. We multiply the send amount by 1+r to get an "exact SPROI" definition. To be as generous as possible to this notion, and to account for the fact that experimental participants typically have a well-known predilection to state whole-number values, we then calculate the least integer greater than this exact value, denoted by ceiling("exact SPROI"). For each send amount, *s*, We label as SPROI all cheating thresholds falling within the interval with integer end-points: [s, ceiling("exact SPROI")]. [5] "Inequality Aversion" refers to a cheating notion which requires equal monetary outcomes, and we label a cheating notion as consistent with inequality aversion if it lies within the smallest closed interval with integer endpoints containing this outcome. As an example, consider s = 1. The total surplus in this case is 10.50 – 1 + 8.05 = 17.55, and half of this surplus is 8.775. Any cheating notion in the interval [8, 9] would therefore be labeled as consistent with inequality aversion. [6] An "Equal-split" (ES) cheating notion entails a cheating threshold of half of the entire amount allocated to the receiver. As with SPROI and Inequality Aversion above, to account for participants' predilection for whole numbers, the definition of ES for each send amount, s, includes all cheating thresholds falling within the smallest interval with whole-number end-points containing a precisely-equal split of the receivers' total earnings: i.e., $\frac{f(s)}{2} \in [n, n+1]$. For example, if a sender sends s = 3, a receiver receives f(s) = 11.30, and $\frac{f(s)}{2} = 5.65$. Consequently, ES for s = 3 would include all cheating thresholds within the interval [5, 6].

Table A5: Proportion of participants whose beliefs about others' cheating notions are consistent with various definitions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
WPROI	0.28	0.22	0.17	0.17	0.25	0.17	0.14	0.17	0.14	0.19
	(0.08)	(0.07)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)
SPROI	0.39	0.36	0.39	0.28	0.31	0.28	0.28	0.28	0.22	0.33
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.08)
Inequality Aversion	0	0	0	0	0.06	0.08	0	0.14	0.22	0.31
	--	--	--	--	(0.04)	(0.05)	--	(0.06)	(0.07)	(0.08)
ES	0.25	0.31	0.33	0.14	0.33	0.36	0.36	0.25	0.22	0.31
	(0.07)	(0.08)	(0.08)	(0.06)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)
Obs	36	36	36	36	36	36	36	36	36	36

Table A6: Beliefs about the amount receivers will return as a function of own cheating notions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Own cheating notion	0.75***	0.75***	0.79***	0.53***	0.57***	0.51***	0.62***	0.84***	0.53***	0.64***
	(0.10)	(0.10)	(0.08)	(0.11)	(0.09)	(0.11)	(0.19)	(0.24)	(0.18)	(0.18)
Constant	0.17	0.48	0.74	2.01**	2.21***	2.74**	2.47	0.99	3.94*	3.04
	(0.30)	(0.46)	(0.50)	(0.78)	(0.78)	(1.04)	(1.79)	(2.26)	(2.02)	(2.13)
Observations	26	31	30	33	34	34	35	35	35	35
R-squared	0.72	0.65	0.76	0.42	0.53	0.42	0.24	0.28	0.21	0.27

Table A7: Beliefs about senders' beliefs about amount receivers will return (second-order beliefs), as a function of beliefs about others' cheating notions, in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Est. others' cheating notion	0.62***	0.61***	0.62***	0.57**	0.59***	0.63***	0.68***	0.64***	0.65***	0.67***
	(0.18)	(0.20)	(0.21)	(0.24)	(0.21)	(0.20)	(0.20)	(0.17)	(0.17)	(0.17)
Constant	0.85	1.40*	1.67	2.42	2.89*	3.12*	2.81	3.59*	3.64*	3.80*
	(0.54)	(0.81)	(1.12)	(1.51)	(1.50)	(1.66)	(1.83)	(1.80)	(1.95)	(2.09)
Observations	36	36	36	36	36	36	36	36	36	36
R-squared	0.25	0.22	0.20	0.14	0.19	0.23	0.26	0.28	0.29	0.30

Table A8: Intentional cheating (reduced form), in-lab sessions

	Sent Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Own cheating notion	-0.78*	-0.35**	-0.08	-0.05	0.02	-0.08	-0.05	-0.02	-0.25**	-0.13*
	(0.43)	(0.17)	(0.12)	(0.11)	(0.07)	(0.09)	(0.11)	(0.11)	(0.12)	(0.07)
Estimate of others' cheating notions	1.08**	0.32	0.15	0.16	0.18	0.05	0.25**	0.11	0.28*	0.34***
	(0.50)	(0.20)	(0.17)	(0.14)	(0.13)	(0.13)	(0.11)	(0.13)	(0.16)	(0.11)
Constant	-0.68	0.27	-0.36	-0.73	-1.13	0.30	-1.49	-0.41	-0.11	-1.98*
	(0.64)	(0.62)	(0.74)	(0.87)	(0.94)	(0.85)	(1.09)	(1.06)	(1.01)	(1.11)
Obs	26	31	30	33	34	34	35	35	35	35

Notes: [1] Each column presents estimates from a Probit model, with the (binary) dependent variable being "receiver intentionally cheats if sent relevant amount." Intentional cheating is defined by sending back strictly less than the receiver estimated senders needed back in order to not feel cheated. This threshold amount is also inserted as a control in each estimate by the variable "Estimate of others' cheating notion." [3] Robust standard errors, clustered by session, in parentheses. *** = significant at 1%, ** = significant ay 5%, * = significant at 10%.

Table A9: Intentional cheating (reduced form), in-lab sessions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Conditional on not cheating (euros returned >= estimated others' cheating notion)										
Est. others' cheating notion	1.24***	1.09***	0.90***	1.08***	1.02**	1.27***	0.44	1.00***	0.93***	0.76
	(0.32)	(0.32)	(0.26)	(0.18)	(0.34)	(0.32)	(0.60)	(0.27)	(0.24)	(0.48)
Constant	0.30	0.51	1.37	1.05	1.36	-0.42	6.27	1.95	2.12	4.74
	(0.77)	(1.16)	(1.32)	(1.07)	(2.28)	(2.61)	(4.90)	(2.58)	(2.53)	(5.05)
Obs	15	16	17	19	15	18	14	11	15	14
R-squared	0.53	0.46	0.44	0.68	0.40	0.50	0.04	0.60	0.55	0.17
Conditional on cheating (euros returned < estimated others' cheating notion)										
Est. others' cheating notion	0.44**	0.43**	0.20	0.37	0.45	0.25	0.44*	0.71***	0.53**	0.40
	(0.19)	(0.18)	(0.21)	(0.22)	(0.28)	(0.22)	(0.21)	(0.21)	(0.23)	(0.28)
Constant	-0.35	0.10	1.84	1.47	0.85	2.84	1.96	-0.19	1.61	2.63
	(0.61)	(0.81)	(1.13)	(1.42)	(2.16)	(1.86)	(2.13)	(2.28)	(2.70)	(3.53)
Obs	21	20	19	17	21	18	22	25	21	22
	0.22	0.24	0.05	0.16	0.12	0.07	0.17	0.32	0.22	0.09

Notes: [1] Standard errors in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents a simple OLS regression of return amount conditional on beliefs about others' cheating notion for the send amount listed in the column heading. [3] The top panel is restricted to observations not involving intentional cheating, while the bottom panel is restricted to observations involving intentional cheating.

Table A10: Send amount (Tobit), in-lab sessions

	Dependent variable = send amount		
	(1)	(2)	(3)
Expected probability of not being cheated	4.29*	4.94**	6.97***
	(2.19)	(2.25)	(1.90)
Expected return from trusting	1.54*	1.57**	1.41*
	(0.81)	(0.75)	(0.77)
Male		1.53*	0.93
		(0.81)	(0.75)
Age		-0.16*	-0.30**
		(0.09)	(0.11)
Math score		-0.34	0.07
		(0.42)	(0.37)
Risk aversion			-0.47**
			(0.17)
Altruism			0.04
			(0.21)
30≤ Income <45			-1.48
			(0.98)
45≤ Income <70			0.03
			(1.10)
45≤ Income <70			1.76
			(1.47)
Income ≥120			-2.99**
			(1.26)
Constant	0.86	5.85	8.66*
	(1.52)	(4.60)	(5.01)
Obs	36	34	32

Notes: [1] Robust standard errors in parentheses. [2] *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [3] Each column presents a Tobit model estimate where the dependent variable is *how much* the sender sends and censoring below 0 is taken into account. [5] "Expected probability of not being cheated" is our measure of participants' subjective beliefs about not being cheated, described in the text. [6] "Expected return from trusting" is the participant's estimate of the proportion of money *sent* that receivers will return, averaged over all 10 possible send amounts. [7] "Risk aversion" is an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism in a separate, unrelated, experiment. This variable takes values from 1 (risk loving) to 10 (very risk averse). [8] Altruism is how much emphasis participants' parents placed on the value "help others" during their upbringing. [9] Income variables refer to (self-reported) annual family income from all sources, in thousands of euros, net of taxes. The lowest category is excluded: "below 30 thousand euros".

Table A11: Proportion of receivers specifying a particular rationale

	Overall	High fee sessions	Low fee sessions
Sender vulnerability	0.42	0.40	0.45
	(0.04)	(0.05)	(0.06)
Positive reciprocity	0.29	0.31	0.27
	(0.04)	(0.05)	(0.05)
Negative reciprocity	0.06	0.06	0.05
	(0.02)	(0.03)	(0.03)
No motive	0.06	0.05	0.08
	(0.02)	(0.02)	(0.03)
Obs	170	93	77

Notes: [1] Raw proportions reported; [2] Standard errors in parentheses; [3] Proportions in each column sum to less than one, with the unaccounted for observations being participants who elected to supply their own rationale rather than one of the four pre-programmed rationale; these self-supplied rationale varied widely and are not easily classifiable.

Table A12: Robustness checks on beliefs, main study data

Panel A: checking for ex-post rationalization

Dependent variable = Expected others' cheating notion

Return amount	Amount sent	High belief pay	(High belief pay) X (Return amt)	Cons	Obs	Individuals	R^2
0.11***	0.85***	0.12	0.05*	1.58***	4254	428	0.5
(0.02)	(0.03)	(0.24)	(0.03)	(0.20)			

Panel B: checking for hedging motives in beliefs

Dependent variable = Expected return amount

Amount sent	Chosen send amount	Risk aversion	(Chosen send amt) X (Risk aversion)	Cons	Obs	Individuals	R^2
0.82***	0.29*	-0.00	-0.02	1.61***	4146	417	0.34
(0.02)	(0.17)	(0.04)	(0.03)	(0.23)			

Notes: [1] Both the top and bottom panel report individual random effects regressions pooling observations across all send amounts. [2] Robust standard errors, clustered by session, appear in parentheses. [3] "High belief pay" is a dummy taking the value of one if the session involved a 20 euro maximum belief pay, and 0 if the maximum possible belief pay was 5 euros; "Chosen send amount" is a dummy variable indicating the amount a participant actually chose to send in the role of sender; "Risk aversion" is an incentive-compatible index of risk aversion obtained from a previous experiment. [4] We drop observations for which we have no measure of risk aversion.

Table A13: Robustness check on own cheating notion, main study data

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Expected return amount	0.66^{A}	0.66^{A}	0.68^{A}	0.58^{A}	0.48^{A}	0.54^{A}	0.57^{A}	0.50^{A}	0.54^{A}	0.52^{A}
	(0.06)	(0.09)	(0.08)	(0.10)	(0.08)	(0.11)	(0.07)	(0.09)	(0.08)	(0.07)
Chosen send amount	-0.18	0.42	-0.30	1.30	-2.22^{B}	-1.16	0.73	1.02	-1.62	1.04
	(0.75)	(0.79)	(0.88)	(1.07)	(0.69)	(1.21)	(0.97)	(2.46)	(1.52)	(2.29)
Chosen send amount X Expected return amount	0.11	-0.32	0.18	-0.29	0.31^{B}	0.07	0.00	-0.02	-0.03	-0.08
	(0.29)	(0.32)	(0.25)	(0.23)	(0.09)	(0.18)	(0.08)	(0.35)	(0.11)	(0.18)
Low Fee	-0.08	0.00	-0.11	-0.26	-0.10	-0.06	0.41	0.23	0.43^{C}	0.05
	(0.15)	(0.18)	(0.11)	(0.20)	(0.24)	(0.13)	(0.27)	(0.27)	(0.22)	(0.26)
Constant	1.99C	2.70^{B}	2.19^{C}	3.89^{B}	6.36^{B}	4.92^{B}	2.55	6.74^{B}	4.92^{C}	5.68^{B}
	(0.96)	(0.91)	(0.93)	(1.48)	(2.09)	(1.81)	(2.01)	(2.17)	(2.23)	(1.97)
Demographic controls?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	311	318	320	328	332	333	334	331	329	329
R-squared	0.37	0.33	0.39	0.30	0.25	0.30	0.32	0.25	0.31	0.29

Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' personal cheating notions. [2] Robust standard errors, clustered by session, appear in parentheses. [3] Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable, "Expected return amount" is a participant's belief about how much a receiver will return for the send amount indicated in the column heading; "Chosen send amount" is a dummy variable indicating the participant actually chose to send the amount in the column heading in the role of sender. [5] Demographic controls are included but not reported for readability. The set of demographic controls is identical to the set reported in Table 6 in the manuscript. "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros. [6] Observations vary over columns because we do not have demographics for all participants and because not all participants reported a cheating notion for all send amounts.

Table A14: Beliefs about the amount receivers will return as a function of own cheating notions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Own cheating notion	0.61^{A}	0.58^{A}	0.52^{A}	0.46^{A}	0.36^{A}	0.46^{A}	0.57^{A}	0.50^{A}	0.51^{A}	0.52^{A}
	(0.06)	(0.02)	(0.02)	(0.06)	(0.06)	(0.04)	(0.03)	(0.10)	(0.10)	(0.10)
Male	-0.30^{C}	-0.49^{B}	-0.34^{C}	-0.53^{A}	-0.43^{A}	-0.37^{C}	-0.23	-0.22	-0.49	-0.28
	(0.15)	(0.17)	(0.17)	(0.13)	(0.10)	(0.19)	(0.22)	(0.32)	(0.31)	(0.25)
Age	-0.01	0.00	-0.01	-0.02	0.01	-0.02	-0.02	0.04	-0.00	-0.00
	(0.01)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.05)	(0.05)	(0.06)	(0.07)
Math score	-0.04	-0.08^{C}	-0.06	0.01	0.10	0.02	0.01	0.05	-0.01	-0.04
	(0.05)	(0.03)	(0.05)	(0.07)	(0.09)	(0.10)	(0.07)	(0.15)	(0.13)	(0.17)
Risk aversion	0.03	0.02	0.04	0.02	0.02	0.07	0.01	0.11	0.06	0.17
	(0.05)	(0.04)	(0.05)	(0.06)	(0.06)	(0.09)	(0.10)	(0.11)	(0.10)	(0.15)
30≤ Inc <45	0.18	0.47^{B}	0.34	0.62^{B}	0.12	0.07	-0.06	-0.31	-0.08	-0.17
	(0.19)	(0.15)	(0.24)	(0.25)	(0.30)	(0.37)	(0.32)	(0.48)	(0.54)	(0.66)
45≤ Inc<70	0.24	0.31	0.32	0.57^{B}	0.29	0.38^{C}	-0.04	-0.25	-0.15	-0.19
	(0.13)	(0.25)	(0.29)	(0.24)	(0.26)	(0.19)	(0.43)	(0.32)	(0.34)	(0.37)
70≤ Inc <120	-0.03	0.17	0.39	0.66^{B}	0.34	0.52	-0.14	0.06	0.01	-0.26
	(0.19)	(0.18)	(0.25)	(0.27)	(0.31)	(0.46)	(0.68)	(0.68)	(0.74)	(0.91)
Inc ≥120	0.26	0.19	0.14	0.39	0.01	-0.36	-0.08	-0.18	-0.42	-0.83
	(0.26)	(0.21)	(0.21)	(0.30)	(0.38)	(0.53)	(0.67)	(0.59)	(0.64)	(0.85)
Low Fee	-0.13	-0.23	-0.30^{B}	-0.14	-0.20	-0.30^{B}	-0.52^{B}	-0.48^{C}	-0.61^{B}	-0.25
	(0.11)	(0.21)	(0.12)	(0.22)	(0.14)	(0.12)	(0.20)	(0.24)	(0.23)	(0.26)
High belief Incentives	0.22	0.71^{A}	0.21	0.09	-0.49	-0.35	0.62	0.08	-0.41	-0.71
	(0.15)	(0.16)	(0.32)	(0.74)	(0.81)	(0.87)	(0.46)	(1.31)	(1.31)	(1.48)
Own cheating notion X High belief Incentives	-0.12	-0.16^{B}	0.00	0.00	0.08	0.05	-0.06	-0.02	0.04	0.05
	(0.06)	(0.06)	(0.05)	(0.11)	(0.09)	(0.09)	(0.06)	(0.12)	(0.12)	(0.11)
Constant	1.00	1.27	1.90^{C}	2.45^{C}	2.24^{C}	2.96^{C}	3.00^{B}	1.70	3.72^{C}	3.44
	(0.81)	(0.74)	(0.84)	(1.05)	(1.06)	(1.30)	(1.16)	(1.36)	(1.71)	(1.84)
Observations	311	318	320	328	332	333	334	331	329	329
R-squared	0.37	0.34	0.38	0.28	0.23	0.28	0.31	0.25	0.30	0.29

Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' beliefs about the amount receivers will return. [2] Robust standard errors, clustered by session, appear in parentheses. Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable is a participant's own cheating notion. Additional demographic controls include: "Math score" = self-reported score on required math exams taken during the final year of high school in Italy; "Risk aversion" = an index increasing in risk aversion obtained from an incentive compatible elicitation mechanism from a prior, unrelated, experiment, which takes values from 1 (risk loving) to 10 (very risk averse); "Inc" = self-reported annual family income from all sources, in thousands of euros, net of taxes. [5] Controls for experimental features are: "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros; "High belief incentives" = an indicator taking the value of one if the session featured a 20 euro payment for an exactly correct belief, and zero exactly correct beliefs paid only 5 euros. [6] Observations vary over columns because not all participants reported a cheating notion for every send amount and because we do not have demographics for all participants. [7] The coefficients and significance levels on the main explanatory variable, "Own cheating notion," are virtually identical if demographics are omitted. From s = 1, …, 10, the coefficients and significance levels are: 0.59^{A}, 0.59^{A}, 0.54^{A}, 0.46^{A}, 0.37^{A}, 0.45^{A}, 0.58^{A}, 0.49^{A}, 0.50^{A}, 0.51^{A}. Moreover, as here, the effect of high belief pay or its interaction with own cheating notion is significant at the 5% level for only one send amount: s = 2.

Table A15: Beliefs about senders' beliefs about amount receivers will return, as a function of beliefs about others' cheating notions

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Est. others' cheating notion	0.83^{A}	0.66^{A}	0.69^{B}	0.84^{A}	0.58^{A}	0.65^{A}	0.51^{B}	0.34^{A}	0.46^{A}	0.45^{A}
	(0.12)	(0.14)	(0.21)	(0.07)	(0.12)	(0.08)	(0.18)	(0.08)	(0.08)	(0.06)
Male	-0.26^{B}	-0.54^{A}	-0.57^{A}	-0.55^{B}	-0.56^{B}	-0.64^{A}	-0.77^{A}	-0.73^{A}	-0.92^{B}	-0.96^{B}
	(0.09)	(0.08)	(0.12)	(0.16)	(0.18)	(0.13)	(0.17)	(0.16)	(0.37)	(0.31)
Age	0.05^{C}	0.05^{B}	0.07^{B}	0.07^{B}	0.07^{B}	0.08^{B}	0.10^{C}	0.09^{C}	0.08	0.06
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)
Math score	-0.10^{B}	-0.13^{B}	-0.09^{C}	-0.18^{C}	-0.06	-0.03	-0.11	-0.19C	-0.07	-0.04
	(0.04)	(0.04)	(0.04)	(0.08)	(0.07)	(0.08)	(0.10)	(0.08)	(0.13)	(0.12)
Risk aversion	-0.01	-0.02	-0.00	-0.00	-0.02	-0.03	-0.04	0.03	-0.00	0.03
	(0.02)	(0.03)	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)	(0.07)	(0.06)	(0.08)
30≤ Inc <45	-0.28	-0.36^{B}	-0.16	-0.39^{C}	-0.11	-0.28	-0.19	-0.14	-0.14	-0.61
	(0.17)	(0.14)	(0.17)	(0.20)	(0.20)	(0.15)	(0.32)	(0.30)	(0.36)	(0.33)
45≤ Inc<70	0.06	0.04	0.26	0.35	0.52C	0.31	0.23	0.25	0.31	0.16
	(0.09)	(0.29)	(0.26)	(0.26)	(0.26)	(0.37)	(0.23)	(0.29)	(0.38)	(0.46)
70≤ Inc <120	-0.15	-0.30	-0.04	-0.02	0.05	0.08	0.02	-0.14	0.14	0.09
	(0.09)	(0.31)	(0.25)	(0.31)	(0.29)	(0.37)	(0.33)	(0.45)	(0.48)	(0.65)
Inc ≥120	-0.17	-0.65^{C}	-0.83	-0.72	-0.81	-0.57	-0.45	-1.08	-0.55	-0.64
	(0.18)	(0.30)	(0.62)	(0.62)	(0.82)	(0.89)	(0.87)	(0.93)	(0.99)	(0.99)
Low Fee	0.03	-0.14	-0.24	-0.23	-0.18	0.02	-0.14	-0.21	-0.14	-0.24
	(0.07)	(0.16)	(0.20)	(0.19)	(0.23)	(0.09)	(0.21)	(0.25)	(0.28)	(0.31)
High Belief Incentives	0.21	-0.20	-0.14	0.43	-0.88	-0.45	-1.57	-4.01^{A}	-2.53	-2.54^{C}
	(0.46)	(0.58)	(1.08)	(0.54)	(1.06)	(0.78)	(1.73)	(0.95)	(1.37)	(1.12)
Est. others' cheating notion X High Belief Incentives	-0.20	0.00	-0.07	-0.20^{C}	0.05	-0.02	0.11	0.34^{A}	0.19	0.19^{C}
	(0.14)	(0.14)	(0.21)	(0.09)	(0.15)	(0.09)	(0.19)	(0.09)	(0.11)	(0.08)
Constant	0.59	1.67	1.28	1.63	2.35	1.76	3.71	6.34^{A}	4.91^{C}	5.57^{B}
	(0.86)	(1.15)	(1.66)	(1.45)	(1.78)	(1.71)	(2.55)	(1.68)	(2.11)	(1.99)
Observations	375	375	375	375	375	375	375	375	375	375
R-squared	0.45	0.40	0.34	0.39	0.33	0.34	0.32	0.32	0.32	0.34

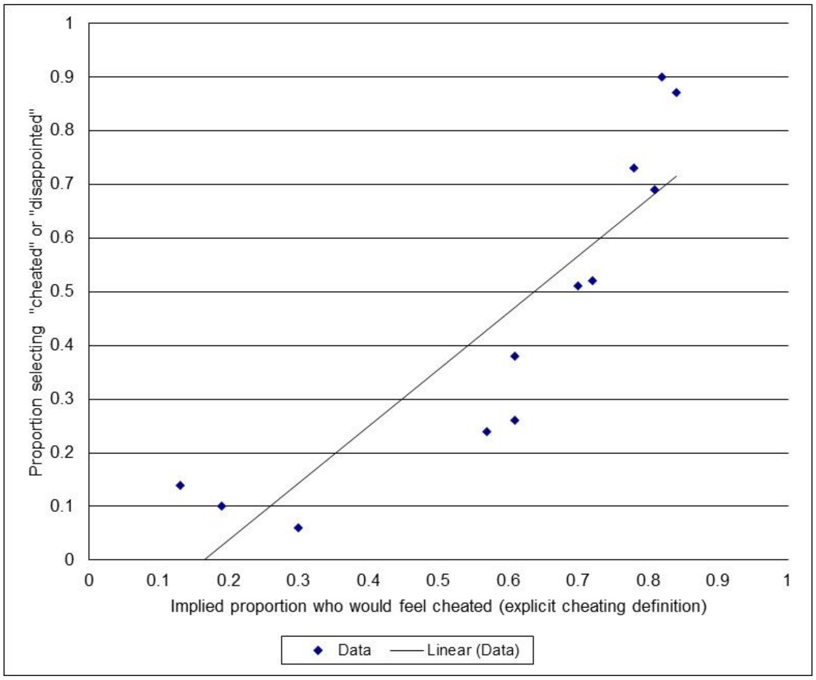
Notes: [1] Each column presents an OLS estimate using as the dependent variable participants' (second-order) beliefs: how much they believe senders believe receivers will return. [2] Robust standard errors, clustered by session, appear in parentheses. Significance levels are denoted by superscripts: "A" = significant at 1%; "B" = significant at 5%; "C" = significant at 10%. [4] The main explanatory variable, "Est. others' cheating notion" is a participant's belief about others' cheating notions. Other demographic controls are identical to those in Table 6, above. [5] Controls for experimental features are: "Low Fee" = an indicator taking the value of one if the session *did not* feature a sending fee of 0.50 euros; "High belief incentives" = an indicator taking the value of one if the session featured a 20 euro payment for an exactly correct belief, and zero exactly correct beliefs paid only 5 euros. [6] Observations vary over columns because we do not have demographics for all participants. [7] If demographics are omitted, the coefficients and significance levels on the main explanatory variable, "Est. others' cheating notion," are virtually identical. From s = 1, …, 10, the coefficients and significance levels are: 0.84^{A}, 0.69^{A}, 0.73^{A}, 0.83^{A}, 0.63^{A}, 0.68^{A}, 0.55^{A}, 0.40^{A}, 0.49^{A}, 0.48^{A}. Moreover, as here, the effect of high belief pay or its interaction with own cheating notion is significant at the 5% level for only one send amount: s = 8.

Table A16: Predicting receiver behavior: second-order beliefs or cheating notion beliefs?

	Send Amount									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Est. others' cheating notion	0.25**	0.32***	0.28**	0.16*	0.30**	0.26***	0.20*	0.29***	0.29***	0.31***
	(0.09)	(0.08)	(0.08)	(0.08)	(0.10)	(0.07)	(0.09)	(0.05)	(0.08)	(0.08)
Second-order beliefs	0.33**	0.12	0.18*	0.20*	0.16	0.12	0.20*	0.11	0.18*	0.09
	(0.10)	(0.09)	(0.08)	(0.09)	(0.14)	(0.09)	(0.10)	(0.07)	(0.08)	(0.10)
Constant	0.00	0.50	2.63**	3.97***	2.03	2.06**	3.62***	4.87**	3.15*	3.42
	(1.45)	(0.85)	(0.93)	(0.76)	(1.13)	(0.86)	(0.65)	(1.45)	(1.39)	(1.93)
Demographics?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	375	374	373	373	373	373	373	373	372	372
R-squared	0.20	0.14	0.14	0.11	0.17	0.12	0.13	0.13	0.16	0.13

Notes: [1] Standard errors in parentheses. *** = significant at 1%, ** = significant at 5%, * = significant at 10%. [2] Each column presents an OLS estimate where the dependent variable in column i is the amount a participant will send back if the sender sends i euros, i=1,…,10. [3] The reported independent variables in column *i* are: "Est others' cheating notion" is each participant's estimate of the minimum amount of money a sender would need back in order to not feel cheated when the sender sends *i* euros, i=1,…,10; "Second-order beliefs" is each participant's belief about the amount of money the sender believes he/she will receive back, on average, when sending i=1,…,10 euros. [4] Each estimate includes our standard set of demographic controls, omitted for readability from the table. These controls are: gender, age, math score, family income and risk aversion. Nothing changes qualitatively if demographics are not included as controls.

Figure A1: Comparison of proportion feeling cheated by elicitation method



Appendix II: Experiment Instructions

In this experiment, you will be randomly paired with another participant and assigned randomly one of two roles: A or B. This pairing will be anonymous. Neither the person in the role of A nor the person in the role of B wil know with whom they have been paired.

The role of A

The player in the role of A is given 10.50 euros and must decide whether to send some all or none of this money to the player in the role of B, the person with whom A has been paired. [If A decides to send some of the this money, A will be charged a fee of 0.50 euros.] For every euro that A sends, B will receive more than 1 euro according to the table below.

If A sends €	1	2	3	4	5	6	7	8	9	10
B receives €	8.05	11.3	13.85	16.05	17.9	19.6	21.2	22.65	24.05	25.3

The role of B

After A makes his or her decision about how much to send to B, B decides how much of the money he or she receives—the amounts in the table above (8.05 euros, 11.30 euros, etc.)—to return to A. The player in the role of B will specify an amount to return for each possible amount they could receive. For example, if A sends 4 euros and B therefore receives 16.05 euros, B must decide how much of this 16.05 euros to return to A; and a decision must be made for every amount A could send (1,2,3,. . . ,10 euros).

Your earnings

For every pair of participants, one in the role of A and one in the role of B, the decisions that both A and B make determine the pairs earnings. Both A and B will be informed of the outcome determined by their choice.

In general:

- If A sends a positive amount to B:

  1. A's earnings will be: € 10.50 – (euros sent to B) + (euros returned by B) – (€ 0.50 fee)
  2. B's earnings will be: (euros received by B according to the table above) – (euros returned to A)

- If A sends nothing to B:

  1. A's earnings will be € 10.50
  2. B's earnings will be € 0.

Specifically, for every pair of players the result of this situation will be determined as follows:

i Every participant specifies their decision for each possible role (A and B).

ii The computer will randomly assign a role to each participant and randomly and anonymously pair each participant assigned the role of A with a participant assigned the role of B.

iii Within each pair, A's decisions will be combined with B's decision to determine the outcome for both A and B.

A Experiment Screens

A.1 Sender decision screen 1

If you are assigned the role of A, do you want to send money to B? If you send money, you will be charged a € 0.50 fee.

Choose "send" or "don't send" on this screen. If you choose "send", you will specify the amount to send on the next screen.

__ Send money
__Don't send money

A.2 Sender decision screen 2

How much money do you want to send if you are assigned the role of A?

__ € 1
__ € 2
. . .
__ € 10

A.3 Receiver decision screens

[There are 10 separate screens. A representative question is below.]

Imagine that you have been assigned the role of B . . .
How much will you send back to A if A sends € 7 and you therefore receive € 21.20?

A.4 Cheating definition screen

If you are assigned the role of A, what is the minimum amount you would need to receive back from B in order to not feel cheated?

If you send €1 and therefore B receives €8.05, you would need back : ____

Insert a number above, or select one of the two following options:
__ This has nothing to do with cheating

__ I do not know

...

If you send €10 and therefore B receives €25.30, you would need back : ____

Insert a number above, or select one of the two following options:
__ This has nothing to do with cheating
__ I do not know

A.5 Belief elicitation

A.5.1 Instructions, screen 1

Now, we begin a new section. In this section as in the previous section, each question can contribute to your potential earnings.

Specifically, in this section you will be asked to estimate the choices other participants made in the previous section. Every question is about the choices of other participants, so please exclude your own actions from your estimations. The accuracy of your estimates will be calculated excluding your own actions as well.

Your earnings from this section will be determined by choosing one of your estimations at random and paying you according to the accuracy of this randomly chosen estimation. Every estimate has the same chance of being chosen by the computer. Your potential earnings from this experiment will be the sum of your earnings in this section and in the previous section.

The formula used to calculate your earnings from the randomly-chosen estimate is detailed on the next page.

A.5.2 Belief compensation formula screen

The method used to calculate your earnings from your estimates is detailed below. The most important thing to notice is that more accurate estimates have higher chances of earning money.

- Your estimate, R, is inserted into the following formula where "r" stands for the true value of the thing being estimated and "r_{max}" is the maximum value this true value can attain.

$1 - \left(\frac{R-r}{r_{\max}}\right)$

- This produces a number between 0 and 1. Call this number "z".
- The computer chooses a number between 0 and 1 with each number in between 0 and 1 being equally likely. Call this number "y".

- If $y \leq z$, you will earn €5.00 [€20.00] for your estimate.
- If $y > z$, you will earn €0.00 for your estimate.

An example

Suppose you are asked to estimate the average amount participants in the role of A send in the previous section of this experiment. And, imagine that this average turns out to actually be €4.00. The maximum value this average could have taken is €10. Therefore "r_{max}" in the equation above is 10 and r is 4. The equation therefore becomes:
$1 - \left(\frac{R-4}{10}\right)$
Notice that the closer your estimate, R, is to the actual value of 4 in our hypothetical example, the larger is z and therefore the larger is the probability of earning €5 [€20.00] for your estimate rather than €0.

- If your estimate is exactly correct, then (R-4)/10 = 0 and therefore z=1. Because the number chosen by the computer is at most one, an exactly correct estimate always pays €5 [€20.00].
- On the other hand, the probability with which your estimate earns you €5 [€20.00] diminishes the farther away from the true value your estimate is: z becomes smaller and so does the chances that $y < z$.

Click continue to begin start the estimation section

A.5.3 Beliefs elicitation screen 1

How much, on average, will players in the role of A send to B's? Insert a number between 0.00 and 10.00 : ___

A.5.4 Beliefs elicitation screen 2

How much, on average, will B's return to A's?

If A sends €1 and B therefore receives €8.05, B's will return on average: ___
...
If A sends €10 and B therefore receives €25.30, B's will return on average: ___

A.5.5 Beliefs elicitation screen 3

What is the minimum amount (on average) that A's will need back from B's in order to not feel cheated?

If A sends €1 and B therefore receives €8.05, to not feel cheated A will need back from B at least: ___

. . .

If A sends €10 and B therefore receives €25.30, to not feel cheated A will need back from B at least: ___

A.5.6 Beliefs elicitation screen 4

What percent of participants in the role of B will return enough money to you (if you are assigned the role of A) so that you don't feel cheated?

If you send €1 and B therefore receives €8.05, what percent of B's will return enough so that you don't feel cheated?: ___

. . .

If you send €10 and B therefore receives €25.30, what percent of B's will return enough so that you don't feel cheated?: ___

A.5.7 Beliefs elicitation screen 5

How much money (on average) do other participants in the role of A believe will be returned to them by B's?

If A sends €1 and B therefore receives €8.05, how much money does A believe B will return? ____

. . .

If A sends €10 and B therefore receives €25.30, how much money does A believe B will return? ____